ACOUSTIC-PHONETIC PROCESSING FOR CONTINUOUS SPEECH RECOGNITION

by

Mary O'Kane

A thesis presented to The Australian National University for the degree of Doctor of Philosophy in the Department of Engineering Physics, Research School of Physical Sciences

January 1981

STATEMENT

Apart from Chapter 4, which was the result of a joint effort between Pietro Demichaelis, Renato De Mori, Pietro Laface and myself, all results and methods discussed in this thesis are claimed to be original except where explicit reference is made in the text.

Mary O'Kane

ABSTRACT

An overall philosophy of continuous speech recognition procedures at the acoustic-phonetic level is developed and consideration is given to ways in which phonetic information derived from the acoustic signal could be used to assist at the level of word recognition. The particular model proposed for obtaining a quasi-conventional phonetic transcript from the acoustic waveform is that of a 'foreign phonetician'. Given the current state of psycholinguistics and associated neurosciences, a system cannot feasibly have built into it all the knowledge that a native speaker acquires unconsciously over many years, but it can be given enough knowledge to enable it to decide which parameters ought to be relevant to recognition of various phonemes. Such knowledge approximates that which a phonetician has when trying to decode and transcribe a foreign language. Several algorithms are presented within the framework of the foreign phonetician model. These include fuzzy algorithms for the recognition of the plosive consonants in Italian and in Australian English, for the recognition of vowels in Australian English, and for distinguishing liquid consonants from nasal consonants in Italian. The parameters used for automatic recognition in these algorithms are those which are generally recognized as being crucial for human aural perception.

In conjunction with the development of these algorithms several problems of particular relevance to continuous speech recognition were investigated. These include the problems of speaker normalization, coarticulation, and word boundary phenomena. Finally the implementation of a highly interactive system, operating according to the foreign phonetician model, which accepts acoustic waveforms as input and produces alternative possible strings of phonetic symbols as output is proposed and discussed.

CONTENTS

ACKNOWLEDGEMENTS

The author is very grateful to the following:

Professor Renato De Mori of the Istituto di Scienze dell'Informazione, Università di Torino, Turin, Italy, for inspiring the work described in this thesis, as well as providing research facilities during the author's sojourn in Italy, and continued encouragement since her return to Australia;

Professor Steven Kaneff and Dr Iain Macleod of the Department of Engineering Physics, Australian National University, for supervising this thesis and providing detailed critical comment of it;

Dr Bruce Millar for many invaluable discussions as well as for making the facilities of the Speech Research Laboratory available;

Hiroaki Oasa for giving the author free access to unpublished data, and for assisting with the preparation of this document;

Colleagues working on Speech Research and Artificial Intelligence in Australia and Italy for lively discussions and helpful suggestions;

Friends who acted as experimental subjects and who helped with computing problems;

Christine Stone who typed this thesis;

Peter Kelly who, as well as being a perfect husband, prepared the diagrams for this thesis.

PREFACE

Automatic recognition of continuous speech has proved possible in a limited and well-specified domain in which the vocabulary is limited and the associated semantics and syntax are a restricted sub-set of those of a natural language. An example of this is the Harpy system developed at Carnegie-Mellon University as part of the ARPA project (Lowerre, 1976). The success of the automatic recognition in such a system is largely due to hypothesis generation based initially on crude acoustic-phonetic decoding and then on sophisticated consideration of the probability that the unknown utterance is any one of the many allowable utterances. This *modus operandi* is feasible because of the restricted nature of the vocabulary, syntax, and semantics, and because the acoustic-phonetic component of the system was specifically adapted to each new speaker.

A system such as the Harpy system cannot easily be extended to tackle the general problem of continuous speech recognition. Widespread commercial viability of machines automatically recognizing speech will come only when systems have been developed which are sufficiently robust in design to be able to recognize conversational modes of speech from a wide variety of speakers, without special adaptation to each new speaker and without requiring that the subject matter be too restricted in nature and form. Realization of such a robust system is critically dependent on improvement in automatic acoustic-phonetic decoding techniques. This thesis considers various ways in which such improvement can be effected, and also emphasises the need for the development of automatic continuous speech recognition systems with strong learning capability to provide a framework in which complex phoneme recognition algorithms can be evaluated

in a realistic environment.

In the first chapter there is a discussion of the problems of deriving a phonetic transcription from an acoustic waveform where that waveform represents the speech (unrestricted as to subject-matter and pre-knowledge of the idiosyncracies of the speaker) of a native of some given language. Along with a discussion of the problems involved, the proposition is made that a system which translates acoustic waveform into phonetic transcription could profitably be designed as a model of a phonetician in a foreign land attempting to obtain a transcription of the language spoken there, a language with which he is, to a greater or lesser extent, unfamiliar. In succeeding chapters some of the particular aspects of acoustic problems which have been introduced in Chapter 1 are examined in depth through the medium of experimental studies, and algorithms for recognition of several classes of sounds are presented. From these attempts to define and assess the difficulty of some major problems for automatic continuous speech recognition, as well as to provide solutions to them, it soon becomes clear that the problems are complicated and the solutions to them non-trivial. For this reason it is considered important to propose a system (the FOPHO system described in Chapter 7) which would allow the actual implementation and development of an operational (albeit at a low efficiency, initially) continuous speech recognition system in parallel with and as a tool for the investigation of the problems which make continuous speech recognition so difficult.

It is to be noted that in Chapters 3 and 4 the experimental material is Italian speech while in Chapters 5 and 6 the experimental material is Australian English speech. During the course of my postgraduate studies I was fortunate to be able to work at the Universita di Torino with the

speech recognition group under the direction of Professor Renato De Mori. Thus Chapters 3 and 4 describe work done by the author mainly in Turin. Throughout the thesis I have retained as basic many of the ideas which guide the building of the automatic speech understanding system in Turin. By adopting parts of the framework of an existing system I have been able to investigate some problems in detail without the need to start a discussion of the problem at its global level. Where some piece of work is based on the Turin system or philosophy, I have given copious references.

Chapters 5 and 6 describe the author's work on the recognition of Australian English speech. Prior to this no automatic recognition studies had been carried out on this dialect, although there are several linguistic studies (Bernard, 1967; Burgess, 1968; Oasa, 1980) on Australian vowels which I have used to extend my own data.

The main results to emerge from this study can be summarized as follows:

(1) the incorporation of context-dependent rules into the recognition scheme significantly improves recognition scores for consonants;

(2) satisfactory vowel recognition can be achieved without any special adaptation to new speakers provided that the sex of a speaker is known (or deduced from the pitch of the voice);

(3) frequency differences between male and female voices are a function of the formant number and frequency range of the formants associated with a vowel; and are a function of the frequency ranges of various consonant parameters.

(4) coarticulation and juncture can be treated as independent effects in

continuous speech;

(5) coarticulation effects are strongly marked in conversational speech show differing dependencies on neighbouring sounds to coarticulation effects in connected or isolated speech.

The following is a list of publications and conference papers arising from the work discussed in this thesis.

Published Papers

P. Demichaelis, R. De Mori, P. Laface, and M. O'Kane, 'Computer recognition of stop consonants', Proceedings IEEE Conference on Acoustics, Speech, and Signal Processing, Washington, 1979.

M. O'Kane, 'New approaches to the acoustic-phonetic component of a speech recognition system', Australian Computer Science Communications, vol. 2, pp.69-83, 1980.

Papers Submitted or In Preparation

M. O'Kane, 'Development of a context dependent algorithm to distinguish between the liquid and nasal consonants', Submitted to the IEEE Transactions on Acoustics, Speech, and Signal Processing.

P. Demichaelis, R. De Mori, P. Laface, and M. O'Kane, 'Computer recognition of plosive consonants using contextual information', Submitted to the IEEE Transactions on Acoustics, Speech, and Signal Processing.

M. O'Kane, 'Mechanisms of coarticulation and juncture at word

boundaries', In preparation.

M. O'Kane, 'Automatic recognition of Australian English vowels in continuous speech', In preparation.

M. O'Kane and H. Oasa, 'On vowel normalization according to sex', In preparation.

Papers Presented at Conferences - Abstracts only published

M. O'Kane, 'Linear prediction in speech analysis', Presented at the Australian Institute of Physics Conference, Sydney, 1976.

M. O'Kane, 'The use of prosody in speech recognition', Presented at the First Australian Language and Speech Conference, Melbourne, 1977.

J. B. Millar and M. O'Kane, 'The ILS software package and its application to the analysis of VCV coarticulation', Presented at the Second Australian Language and Speech Conference, Melbourne, 1979.

M. O'Kane, 'Coarticualtion across word boundaries', Presented at the Tenth International Conference on Acoustics, Sydney, 1980.

Chapter 1

INTRODUCTION

1.1 THE TOPIC DISCUSSED

This thesis is concerned with aspects of the design and development of an automatic acoustic-phonetic analysis system for continuous speech, i.e. a system which accepts a speech waveform as input and produces a string of phonetic symbols in word-like groupings as output. Syntactic recognition rules are developed for various classes of phonemes occurring in a variety of contexts. These rules are consistent with known features of human speech perception - indeed it can be argued that this system of rules is a model of human speech perception. The rules are sufficiently robust to be able to detect and categorize phonemes occurring in continuous, conversational speech as produced by adult speakers, male and female, of a given language; such robustness being due to the extensive exploitation of the naturally occurring information redundancy of the speech code and the power of Fuzzy Set Theory when applied to classification tasks.

Although a model of word recognition into which the results of the acoustic-phonetic decoding would fit, is given in this chapter, the acoustic-phonetic work described in subsequent chapters could be used in many Speech Understanding Systems. Indeed, the author's work described in Chapters 3 and 4 is already implemented as part of the University of Turin's Speech Understanding System.

The main aims of the thesis are:

(1) to investigate the general problem of phoneme recognition for the case of continuous speech with a view to discovering why most automatic recognition systems have not achieved very accurate phoneme recognition to date;

(2) to find new ways to incorporate information from linguistics, psychology, and neurophysiology to produce better phoneme recognition rules;

(3) to clarify the quality of phoneme recognition needed for successful interaction with other components of a continuous speech recognition system.

(4) to investigate in detail several specific problems for acoustic-phonetic recognition. These include the development of context-dependent recognition rules, an investigation of the possibility of speaker independent phoneme recognition, and an analysis of what happens at word boundaries.

The work was approached from an Artificial Intelligence viewpoint i.e. with the absolute question in mind: can an automated system be developed which can function like the human brain in that it can hear human speech, decode it and produce a suitable response? However, much of the work in the development of a rule-based system of phoneme classification and recognition is primarily an exercise in computational phonetics and many of the results discovered would seem to have more relevance to that field than to the field of Artificial Intelligence. Nevertheless the fact that the work is approached from an A.I. level results in questions being asked and experiments being designed which would generally not be in the normal scope

of phonetics.

In this chapter the 'Foreign Phonetician' model of an acoustic-phonetic decoder and subsequent word recognizer is introduced. (This is an automatic system which operates like a phonetician working in a foreign land on a language which is unfamiliar to him.) Also there are reviews of research on human speech processing and some aspects of speech production. Such reviews are given to highlight what features an automatic speech recognition system must incorporate if it is to act in a manner consistent with the human speech perception mechanism. The relation of such a system to already existing Speech Recognition Systems is also discussed. Details of the possible implementation of the acoustic-phonetic model introduced in this chapter are described in Chapter 7. In Chapter 2 there is a discussion of Fuzzy Set Theory and its use in syntactic pattern recognition. Chapter 3 describes the implementation of a system to distinguish the liquid consonants from the nasal consonants, and provides a reasonably simple example of the type of fuzzy recognition rules used throughout the thesis. Chapter 4 describes the development of a system of recognizing the individual stop consonants in Italian. Chapter 5 describes Australian English vowel recognition with particular emphasis on speaker-independent vowel recognition. Chapter 6 describes an extensive experimental examination of two crucial problems for continuous speech recognition - the problems of juncture and coarticulation across word boundaries. A recognition algorithm for plosive consonants in Australian English is also given in Chapter 6. Chapter 7 describes the overall acoustic-phonetic model in detail and suggests how it should be implemented. Chapter 8 is a discussion of the achievements of the work described in this thesis.

Recognition algorithms for selected phonemes in two languages are described in this thesis. The work described in Chapters 3 and 4 was mainly carried out while the author was at the Istituto di Scienze dell'Informazione, Turin, Italy, where an Automatic Speech Understanding System is being developed for Italian. Italian, a phonetically pronounced* language with a small number of vowels (5) is particularly amenable to automatic recognition although there is a problem with regional variations. Chapters 5, 6, and 7 are descriptions of recognition algorithms for Australian English. Like all varieties of English, Australian English is not phonetically pronounced and has a large number of vowels (~18). However unlike most national varieties of English it is remarkably stable as regards regional variations (Bernard, 1969; Oasa, 1980).

1.2 THE PROBLEM OF RECOGNIZING CONTINUOUS CONVERSATIONAL SPEECH AS SPOKEN BY ANYONE

Commercial systems are available which recognize small vocabularies of up to fifty words spoken by a limited set of speakers on whose voice parameters the system has been trained. These systems operate on word template matching techniques. If one applies such techniques to the problem of recognizing large vocabularies of say 1000 words one faces a gigantic storage problem - that of storing 1000 word templates for each speaker using the system (Kohonen, 1980). Also use of exclusively template matching techniques will be successful only if the words are spoken separately because when words are spoken in continuous speech or even connected speech (words spoken in a connected string e.g. telephone numbers) it is found that there are coarticulatory interactions between words (Klatt and Stevens, 1973; Klatt, 1979) and also that words will be pronounced differently depending on whether or not they are stressed. In

* Phonetically pronounced with reference to the orthography.

unstressed positions vowels will often be reduced and the schwa vowel /ə/ will often be substituted for the vowel found in the isolated citation form when the word is in a stressed position. Also word endings will often be slurred or left out altogether e.g. the phrase 'Can you understand it' will often come out in conversational speech as:

kənjuʌnəstænət

with both d's in 'understand' not being pronounced at all. Thus it must be concluded that word template matching techniques alone are inadequate for tackling the problem of continuous speech even when spoken by just one speaker.

It would seem that a way of obtaining at least a rough phonetic transcript of what is being said must be found. This might be done by crude template matching or by a more sophisticated system of syntactic phoneme recognition algorithms. The phonetic transcription can then be passed to another level of processing which uses phonological and syntactic information to discover where word boundaries might be and to what grammatical classes various words might belong, e.g. a word ending in the suffix 'ly' will be tagged as a probable adverb. Semantic information is used to test the sense of the phrase under consideration For example if it is suggested that the sentence being analyzed is:

The curl sang a song

semantic knowledge would allow the system to detect this as a nonsensical phrase and the system would probably postulate that a /k/ instead of a /g/ sound had been recognized at the acoustic-phonetic level and thus the sentence should read:

And if one is working in a specified task domain, pragmatics can be used to eliminate unlikely words and phrases. For example the word 'kangaroo' would be given an unlikely rating if the specified task domain was plumbing.

The ARPA speech understanding project (which is reviewed in Section 1.6) demonstrated that use of 'higher level information' i.e. syntax, semantics, and pragmatics, was essential to the achievement of good recognition scores for continuous speech (Medress et al.,1976). Indeed the system that performed best within the ARPA directives, the Harpy system, developed at Carnegie-Mellon University used relatively crude template matching at the acoustic-phonetic level and yet managed because of the power of the higher level information incorporated into it to achieve better than 95% accuracy. This accuracy was also partially due to the fact that Harpy was working in a constrained task domain and allowed training for new speakers to the machine.

What happens when we decide to tackle the problem of continuous speech as spoken by the 'man in the street' using his ordinary conversational voice; and speaking about any topic? This is the problem which must be solved in order to reach the main envisaged use of Speech Recognition Systems as a direct means of communication between a layman and a computer. It has been demonstrated that speech is by far the most efficient human communication medium (Ochsman and Chapnais, 1974).

In the unconstrained everyday situation all levels of the speech recognition process face significantly harder tasks. Because there is no specific domain of interest the use of pragmatics is less rewarding. The

non-fluid nature of continuous discourse (i.e. the effects of hesitations and incomplete sentences) mean that the semantic and syntactic interpreters must be more skilful than they need be for more controlled formal interaction. The job of the acoustic-phonetic analysis system is very much harder as fast conversational speech is rarely characterized by carefully enunciated phonemes. Also the system must be able to handle the idiosyncracies of a large number of speakers. And in this case, the higher levels of processing will be more than ever dependent on the output at the acoustic phonetic level. So rather than merely aim to achieve a reasonable phonetic transcription at the output of the acoustic-phonetic recognizer, a system producing as highly accurate a transcription as possible would seem to be desirable. A system which could be developed towards this goal is proposed in Section 1.6, where the 'Foreign Phonetician' model of acoustic-phonetic recognition is discussed.

In the following two sections, aspects of human speech production and perception mechanisms are considered with a view towards incorporating features of these mechanisms into automatic recognition procedures.

1.3 COARTICULATION

Coarticulation is the phenomenon whereby the manifestation of one sound is affected by the sounds adjacent to it in an utterance. Most commonly, the affected sound will acquire some of the characteristics of its neighbours; thus in the utterance 'spoon' we find that the /p/ and /n/ productions will be lip-rounded due to the presence of the ip-rounded vowel /u/.

The usual explanation for this effect is that it is physiological. Two types of coarticulation are delineated - anticipatory and carryover. Anticipatory coarticulation occurs when, during the production of a sound, the articulators approach their positions for the following sound. An example of this is the lip rounding encountered in the production of /p/ in 'spoon'. Carryover articulation occurs when the articulatory positions for a given sound lie between the positions for the previous sound and the canonical positions of the sound being produced. This form of coarticulation can be seen in the rounding of /n/ in 'spoon'. Anticipatory and Carryover coarticulation are also called 'right-to-left' or 'backward' and 'left-to-right' or 'forward' coarticulation respectively.

The assumption that coarticulation is accounted for by articulators preparing for future sounds or sluggish articulators not moving sufficiently from their positions for previous sounds presumes that we articulate in segments i.e. an utterance is made by productions of the phonetic components. This presumption has been questioned by Hammarberg (1976) who points out that there is no evidence that speech consists of segment produced after segment: rather the natural flow of speech would suggest that the dynamic speech pattern is mind-programmed and thus coarticulation is the result of rules which operate at the intentional level so that the articulators allow their movements to be as smooth and free flowing as possible.

In artificially intelligent programs written to recognise speech very little account has been taken of coarticulation; instead it is usually hoped that coarticulation effects will not give too great a deviation from the 'standard' form of each sound. Nevertheless this approach was often seen to be inadequate (Weinstein, McCandless, Mondshein, and Zue, 1975).

Whether or not coarticulation rules are included in a speech recognition procedure would depend on how much emphasis is placed on recognition at the acoustic-phonetic level, and just how much deviation from the 'standard' phonetic forms is induced by coarticulation. As one aim of this project is to obtain high quality recognition at the acoustic phonetic level it is felt necessary to know just what effects coarticulation has. Certainly when contextual effects were included in recognition schemes (Demichaelis, De Mori, Laface, and O'Kane, 1979) recognition rates were higher.

To emphasize the importance of coarticulation a brief survey of the literature relating coarticulation to speech production and perception is presented. In recent years there has been a series of major articles in the Journal of Phonetics examining the nature of coarticulation (Daniloff and Hammarberg, 1973; Hammarberg, 1976; Kent and Mininfie, 1977; Fowler, 1980). These papers propose various strategies which might be used in the production of normal (i.e. coarticulated) speech. Theories of production and perception must be able to account for all the coarticulatory phenomena. A theory of perception has to be capable of explaining how it is that if asked whether the vowel sound in 'fit' is the same as the vowel sound in 'dig' most people will answer affirmatively, although an examination of the spectra of the two vowels will reveal differences due to the influence of differing neighbouring consonants. It might be objected that the differences seen in the spectrograms are not great enough to be heard. However there is evidence that if the sounds are excised from the two words and interchanged so that listeners are presented with a word in which the /i/ sound from 'fit' is placed between the /d/ and /g/ from 'dig' the word will not sound like 'dig' anymore (Lindblom and Studdert-Kennedy, 1967). From this it can be concluded that coarticulatory effects are automatically allowed for in phoneme recognition. Studies on 'deaf speech'

in which coarticulatory effects are abnormal (Rothman, 1976), consistently show that phonemes in deaf speech sound 'odd' to normal hearers. Thus a perceptual theory must account for coarticulation as a phenomenon which is usually taken for granted but its absence is noted if it is not used.

Theories of speech production must account for the fact that coarticulation can extend over several sounds (Moll and Daniloff, 1971; Benguerel and Cowan, 1974), across word boundaries (Su, Daniloff, and Hammarberg, 1975; O'Kane, 1980), that it can extend in both directions and is speaker dependent (O'Kane, 1980), and it is possibly language dependent (Kent and Mininfie, 1977). Much of the literature on coarticulation effects particularly on effects which spread over several sounds consists of reports of electromyographic experiments on articulator muscles (Gay, 1977; Butcher and Weiher, 1978). Also there is ample spectrographic evidence for coarticulation (Ohman, 1966; Lindblom and Studdert-Kennedy, 1967). This spectrographic evidence indicates that automatic recognition procedures should take coarticulation into account as spectrograms are generally the basic patterns on which recognition algorithms work.

As the interest here is not merely to aim towards a working recognition system but to develop a system which operates like the human brain it is interesting to know what the perceptual correlates of the observed coarticulation production effects are. The literature on this topic is not very extensive and what exists gives incomplete and at times contradictory results. Two types of experiments are used: one in which synthetic speech is used and examples of such speech, in which certain parameters are manipulated to obtain differing listening conditions, are played to a panel of listeners. From such experiments comes indirect evidence that for a phrase to sound normal coarticulatory effects are

necessary (Harris, Hoffman, Liberman, and Delattre, 1958; Lindblom and Studdert-Kennedy, 1967). In the other type of experiment, segments are removed from natural speech and listeners are asked to judge whether or not they can tell anything about the missing sound(s). In the experiment by Lehiste and Schockley (1972) the results were largely negative indicating that little could be deduced about an excised sound by listening to neighbouring sounds. It should be noted that in this experiment the speech of only one speaker was used however. Kuehn and Moll (1972) on the other hand (again using only the speech of one speaker) demonstrated that features of the excised sound could definitely be deduced from listening to neighbouring sounds. The order of perceptibility of features was place of production, voicing, and manner. Similarly Ali, Gallagher, Goldstein, and Daniloff (1971) found that if all the parts of a consonant occurring at the end of a nonsense CVVC sequence were removed listeners could tell whether the consonant was a nasal consonant or not. The diversity of results from these experiments would seem to be due to differing coarticulation strategies being used by the different speakers.

From the above discussion it can be seen that there is ample evidence that coarticulation is an important factor in speech perception. Many of the works cited give displays that illustrate that perceptual coarticulatory effects are very noticeable in spectrograms. So it would seem to be very reasonable to account for coarticulation in a recognition system. But how? One way would be to follow the Wickelgren (1969) model of speech perception which postulates that the brain stores all possible allophones of a language i.e. rather than merely storing a representation of an ideal phoneme X it will store a set of phonemes ${}_{y}X_{z}$ where Y and Z specify the context. This does not seem to be a very efficient way for the brain to operate as the number of sound representations that must be stored

would be extremely large. If this procedure is adopted for automatic recognition then a storage problem would almost certainly result. Also as will be demonstrated in Chapter 6 different people use coarticulation to differing degrees. And one speaker will not always be consistent in his use of some coarticulatory effects. It would seem not unreasonable to postulate that the brain must use a strategy which is alert to the possibility of coarticulation in that it knows which features of sounds are most likely to be coarticulated and to what extent. Doubtless the brain uses coarticulatory effects to anticipate future possible sounds and to check on previous ones. In the automatic recognition system described here most recognition rules are context dependent - the actual strategy will be discussed in later chapters.

1.4 SPEECH PERCEPTION STUDIES AND THEIR IMPLICATIONS FOR AUTOMATIC SPEECH UNDERSTANDING

1.4.1 Phoneme Perception

The well-known speech production effect of coarticulation was discussed in the preceding section. Coarticulation was seen to be an effect that could also be perceived. Here other aspects of the speech perception-by-humans process are considered and again we see, as in the coarticulation case, many of these perception effects such as the perception of formants and the perception of prosody are directly related to mechanisms of speech production. On the other hand, there are effects such as categorical perception and word and syllable perception which do not seem to be directly connected to speech production; rather they seem to have been evolved in order to deal with the noisiness and imprecision which result from human speech production. The implications of these

facets of speech perception for automatic speech-understanding systems will be discussed.

Speech perception is the process which has speech waveform as input and some structure which is linguistically meaningful to the perceiver as output. Thus the speech perception process is different from the process by which we hear and interpret music and other non-speech sounds. Also, when one hears speech in a foreign language which is not well known to the hearer the perception process will not be able to complete its function because when one perceives speech one generally does so with reference to some well-known linguistic framework (one's native language) and uses the accumulated knowledge of this framework at a very early stage in perceptual processing. It would seem that generally our first conscious knowledge of what we have heard comes after the incoming speech has been decoded (sequentially) into words. For example, Savin and Bever (1970) had subjects monitor a list of nonsense syllables in order to detect a target syllable or a target phoneme from that syllable. They found that subjects responded significantly more slowly to the phoneme targets than to the word targets, indicating that the syllable was perceived and then analysed so that its constituent phonemes could be examined. Rubin, Turvey, and Van Gelder (1976), in experiments designed similarly to those of Savin and Bever, except that words as well as non-words were used, asked subjects to detect a target phoneme and found that it was detected more rapidly if it occurred in a real word than if it occurred in a non-word. This has been interpreted as indicating that a search of the lexicon to test whether the word is legal is carried out before the hearer becomes consciously aware of the word. Thus it would appear that we have no conscious access to the phoneme (if such a unit exists). Indeed, the concept of the phoneme would seem to be a learned one, for Morais, Cary, Algeria, and Bertelson (1979)

found that illiterate adults could neither add nor delete a phoneme at the beginning of a non-word, even though the addition or deletion would have resulted in the production of a real word; literate adults who had been exposed to a quasi-phonetic alphabetic writing system (in this case, Portuguese) had no difficulty with the addition/deletion task. Nevertheless Foss and Blank (1980) have shown that subjects can respond to a phoneme prior to word recognition. They found that subjects who were asked to monitor a target phoneme occurring in a word in a sentence context were as fast in detecting the target in a word, whether of high or low usage frequency, or a non-word. Apparent inconsistencies between the results of this experiment and those of Savin and Bever (1970), Foss and Dowell (1971), and Rubin et al (1976)) have been discussed by Foss and Blank and attributed by them to variations in experimental design. Of course, this does not mean that we necessarily consciously perceive phonemes all the time. Rather it indicates that we can, on occasion, become conscious of some smaller-than-syllable sized unit of speech. That it might be the phonemes is indirectly hypothesized by the existence of phonetic spoonerisms.

1.4.2 Word Recognition in Continuous Speech

Leaving aside temporarily the issue of phoneme perception, let us consider the proposition that lexical information is available to the processing system at an early stage. Very convincing evidence for this comes from experiments involving sound replacement and shadowing. Warren (1970) showed that if a phoneme was excised from a word in a sentence and was replaced by a non-speech sound such as a buzz or a cough, the listeners almost invariably perceived the missing phoneme as though it had been actually present and were unable to indicate correctly where the non-speech

sound had occurred. The subjects of Warren's experiment had enough phonetic information to find the right lexical phonological match for the perceived word. Cole (1973), Cole, Jakimik, and Cooper (1980),and Marslen-Wilson and Welsh (1978) have seen the same effect in their experiments where they asked subjects to detect the mispronunciation of an embedded word. Subjects generally did not detect the mispronunciation, but when they did, they succeeded faster if the altered phoneme was towards the end of the word. This occurred because by the time the altered sound was being processed the word had almost certainly been recognized and all that was required was a matching process to see whether mispronunciation was present. In shadowing experiments (where the subject repeats the perceived speech stimulus as fast as possible), Marslen-Wilson and Welsh (1978) found that subjects often fluently restored a mispronunciation to the 'proper' rendering of the word unless the mispronunciation occurred early in the word. If they did detect the mispronunciation they slowed down considerably in their shadowing task.

From a consideration of the shadowing experiment and of mispronunciation detection and target monitoring experiments Marslen-Wilson and Welsh have proposed a model of word recognition in continuous speech. Rejecting both Morton and Long's (1976) logogen model (an interactive model where both lexical information and phonetic information can trigger word recognition) and the Autonomous Search Model of Forster (1976) (a bottom-up model in which word recognition relies on phonetic decoding) on the basis of the results of various response timing experiments, they propose that by the time the first few phonemes of a word have been heard there will have been a multiple parallel activation of an entire class of word candidates (the class of all words whose beginning is consistent with the class of current phonetic hypotheses for the beginning segment). As more of the

input is heard, more and more word elements (all assumed to be actively monitoring the input) will remove themselves from the cohort of possible solutions. The cohort will be further reduced by active on-line constituents of the semantic and syntactic context. Context would also have the effect of indicating with what precision the rest of the input should be analysed once the word has been identified from the first few syllables. If the context effects are compelling perhaps the perception mechanism will not even bother analysing the rest of the word in any detail (this explains the effect of hearing an incorrectly produced sound in a word as the 'correct' sound). However if there is some doubt the rest of the word could be analysed 'just to be sure'.

1.4.3 Implication of Theories of Perception for Automatic Speech Recognition

So what does the discussion above imply for automatic speech recognition? The most obvious thing is that the human brain, the most efficient recognition system we know of, uses high-level lexical and syntactic information at a very early stage in the processing system. This has been recognized for some time and the importance of adopting a similar strategy in artificial speech recognition was highlighted by the success with which the Harpy system (a system relying heavily on high-level information i.e. a top-down system) fulfilled the ARPA Project goals (see next section). But as mentioned earlier, in the general non-restricted conversational speech problem how much top-level information can be made available? Let us consider the possibility of implementing an automatic speech recognition system designed according to the Marslen-Wilson and Welsh speech processing model. First consider the lexicon. This could be a great number of files each one primarily consisting of a key-word.

Typically an adult's word vocabulary is of the order of tens of thousands. Now when a word is produced in conversational context its production will vary with the context as well as with the speaker's dialect, age, and sex. Does a typical file of the lexicon contain lists of all the possible productions of a word i.e. templates of all the possible variations? Probably not, as this would mean that to be able to deal with all possible contexts and speakers an immense amount of storage is needed. Also how would we be able to understand a speaker we had never heard before? So it would seem more sensible to postulate that the recognition system has a bank of phonological rules at its disposal that operate on a phonetic decoding of the acoustic input and act as the mediator between the phonetic data and the lexicon which is then assumed to contain only one representation of each word. Thus for the two inputs:

wɒt də ju wɒnt - What do you want?

wɒt wɪl i du - What will I do?

the phonological rules will be able to deduce from the phonetic representation that both phrases begin with the word 'what'. Also phonological rules would be partly responsible for word boundary detection. Other factors that would be important in detecting word boundaries are phonetic juncture effects (see Chapter 6), prosodic rules, and sound position rules of a language (e.g. a word in English cannot begin with /ŋ/). This business of word boundary detection is important because target monitoring effects have shown that word recognition in a sentence proceeds in a time-sequential manner although limited memory storage (Massaro, 1974) means that some feedback could occur.

So the automatic recognition system will have phonological rules hardwired into it. These rules would serve the function of continuously monitoring the phonetic code to guide the data driven lexical search. Of

course as the list of word candidates is narrowed the lexicon will be increasingly certain of the identity of the word and will suggest hypotheses to the phonological monitor which will verify or reject these after inspection of the phonetic code. (After all, verification is easier than identification.) We are saying that word recognition proceeds by the input initially alerting a set of possible candidates and that these candidates become increasingly possible (or are rejected) and as this happens the phonetic input is no longer used as a selecting device but rather as a checking device. But in implementing such a system we would find that although we are aware of many phonological rules (Chomsky and Halle, 1968) and although we know of many ways of making syntax and semantics aid in the word selection process we soon find that our conscious knowledge of phonological, semantic, and syntactic rules is not nearly as comprehensive as the subconscious knowledge of these rules which is used by the average adult in speech processing.

This is where the Foreign Phonetician model is probably a more realistic solution than using a model of an average adult in his native environment. The foreigner must obtain as good a phonetic transcript as possible (we know that this can be done from Foss and Blank's experiment) because he does not have as good a higher-level structure as the native. Because of the inadequacies of the phonological rules many legal words will be processed more in the manner of non-words i.e. attention will be given to the phonetic code for the entire length of the word. Also feedback mechanisms will have to be more active than they probably are in normal speech processing. Thus a word might be totally re-examined because the foreigner knows that the speaker is not producing non-words although he (the foreigner) has just processed the word as he would a non-word so now he must try to locate possible wrong moves and then alert new lexical

hypotheses. The foreigner is definitely much more inefficient than the native at speech processing. However, most foreigners learn to quickly recognize a limited (survival) vocabulary. Thus a Foreign Phonetician model automatic speech understanding system would be most useful (and lifelike) if it incorporated a learning mechanism i.e. if it had some kind of mistake memory. In practice this could be achieved by analysis of both mistakes and of processes that take a long time to find the right answer. From this analysis suitable modification of the rule structure could be achieved. Thus such an automatic recognition system should be easily modifiable. It also must have a method of evaluating and scoring competing hypotheses such that all likely hypotheses are pursued and all increasingly unlikely ones discarded after a suitable delay time. (Immediate discarding of hypotheses is probably unwise in case later information indicates that they should be re-activated.) In the brain this is very likely done in a fuzzy manner (Massaro and Oden, 1980). That this can easily be implemented in an artificial system is demonstrated later (Section 1.6). Before leaving the Foreign Phonetician model it is worth briefly considering some other ways of capitalizing on it to obtain better recognition. One of these would be to make it interactive (even let it have a foreign accent if it must!), and be able to ask for clarification of a concept or to have a phrase repeated more slowly, it might also repeat what it thinks it heard and ask for verification of the part it is not sure about. Something else that is characteristic of a foreigner is his typically limited vocabulary - so perhaps the automatic system should have only a small vocabulary initially and it can be increased as need demands and as other levels of processing become more efficient. When it encounters a word it does not know, it would have to tell the speaker to re-phrase what was said in a simpler way.

In our survey of human speech perception we have concentrated so far on the processing mechanism after the acoustic waveform has been translated into phonetic code. Now we will review what is known about that initial step of interpretation of the acoustic waveform and the deduction of the phonetic code from it. After all the task of an acoustic-phonetic component of a speech recognition system is to deduce the phonetic code from the acoustic input. The review above of post-phonetic processing gives indications on how an acoustic-phonetic processor has to fit into the overall automatic recognition scheme. What is discussed below is aimed at highlighting which acoustic parameters are phonetically relevant. As we saw above, the earliest level of speech processing that is available for conscious monitoring is the phoneme. Thus evidence on the nature of pre-phonetic processing is indirect and is generally obtained by the use of special experimental paradigms and tools. One such tool is synthetic speech which has been used extensively to uncover which features of a sound are vital to the correct perception of that sound. A good example of this is the work done during the 1950's on the stop consonants at Haskins Laboratories (Liberman, Delattre, and Cooper, 1952). Real speech experiments on the other hand are used to investigate the interaction of phoneme parameters in fluent speech. Motivation for many of the studies on phonetic perceptual processing comes from studies on speech production on the assumption that there could very well be a link between speech production and speech perception. (*Vide* the Motor Theory of Liberman, Cooper, Shankweiler, and Studdert-Kennedy (1967).)

1.4.4 Formant Perception

One known link between speech production and speech perception has to do with formants. A vowel sound is characterized by these prominent spectral resonances. Examination of the speech production reveals that these resonances are directly related to the characteristic resonances of the vocal tract cavities used in the production of that vowel. The relation between these and the dimensions of vowel perception was demonstrated by Pols, van der Kamp, and Plomp (1969) who had subjects measure similarities between a set of pre-recorded vowels by the method of triadic comparisons. A dimensional analysis of the pooled results showed that two dimensions explained 70% of the variance, three dimensions 80%, and four dimensions 90%. These same vowels were analysed using a set of third-octave filters and dimensional analysis was performed on the outputs of these filters. When the vowels were plotted in two-dimensional space representing the two principal dimensions, it was found that plots for the perceptual and physical data were similar. Furthermore, yet another similar plot resulted when the first two formant frequencies were plotted along two dimensions. From this it was concluded that the first two formant positions were critical to vowel perception. There was also some evidence that the nasal and the liquid consonants could be described in the same way.

1.4.5 Front Cavity Resonance

Another recently investigated production-perception correlate concerns the front-cavity resonance. Kuhn (1975) pointed out that in the spectrum of fricative speech (in which palatal frication is substituted for laryngeal voicing and the nasal port is kept closed) the spectrum of the front cavity appears as the most prominent spectral component. A

comparison of spectrograms of sounds produced in both normal and fricated speech illustrated the fact that the front cavity is not always associated with the same numbered formant: for central and back vowels the prominent resonance of fricative speech (Ff) occurs near the position of F2 in normal speech while for high front vowels such as /i/ Ff is associated with F3 in normal speech. All this would seem to indicate that the formant carrying most information about the place of articulation of the vowel is the formant which corresponds to the front-cavity resonance (a result which is not at all surprising when considered from a production point of view). Further experiments by Kuhn (1979) in which Ff was compared with individual formants alone or in selected combination as a means of cueing place of articulation for stop consonants, showed that Ff was the best overall cue to place perception; that the second and third formants of normal speech showed a change in their ability to provide place perception for stop consonants that is in the direction of expected front-cavity affiliations with the third formant doing as well or better than the second formant only for stops before the vowel /i/; that the stop burst seemed to be an important place cue for the alveolar (t,d) and the velar (k,g) stops where it can be thought of as the initial manifestation of the front-cavity resonance. The importance of these discoveries lies in the way in which they indicate that the various cues for stops are actually all manifestations of the same phenomenon - the front-cavity resonance. Also it indicates that F3 will play an important role in place of articulation determination when the sound in question is a high front vowel or a sound coarticulated with a high front vowel.

1.4.6 Categorical Perception

Let us now turn to the effect known as categorical perception. This effect is one which is a special perceptual effect which is perhaps best illustrated by description of an early experiment in which the paradigm was first observed. Liberman, Harris, Hoffman, and Griffith (1957) constructed thirteen different synthetic CV syllables that differed only in the direction and extent of the initial second formant transition into the vowel. The starting frequency of F2 differed in equal frequency intervals over a range sufficient for the perception of the stop consonants, /b/, /d/, and /g/. Played in order the CV syllables give the impression that the changes from /ba/ to /da/ to /ga/ are abrupt. When subjects were presented with a pair of these CV syllables it was found that discrimination between the two presented stimuli was only possible if the two stimuli were drawn from across a phoneme boundary. When stimuli drawn from within the same phoneme boundary were presented it was found that they were always said to be the same. The categorical perception phenomenon has been observed in many studies on several acoustic dimensions. Pisoni (1973) constructed three formant stimuli in which both the second and the third formant transitions were used to cue place of articulation. When presented to experimental subjects these stimuli were perceived categorically. Many studies have shown that voicing in stop consonants is perceived categorically (Liberman, Harris, Kinney, and Lane, 1961; Abramson and Lisker, 1970). Larkey, Wald, and Strange (1979) have reported categorical perception of place of articulation cues for nasals in initial and final positions. Cutting and Rosner (1974) demonstrated categorical perception between fricatives and affricates for the same place of articulation. In all these studies the same paradigmatic effects are seen. Cross-category boundary discrimination tasks are always done accurately

while within category tasks always yield near-chance results.

Not all acoustic dimensions are perceived categorically however. Experiments on acoustic parameters associated with vowels showed that they were not perceived categorically (Fry, Abramson, Eimas, and Liberman, 1962; Pisoni, 1971) but rather exhibited continuous perception. Nevertheless vowels in a final VCV context were found by Stevens (1968) to be perceived more categorically than isolated steady-state vowels. Perhaps categorical perception is basically a feature of our perception of continuous speech.

Models of speech perception have sought to account for the phenomenon of categorical perception. Perhaps the currently most controversial of these models is that which proposes innate opponent process feature detectors (Eimas and Miller, 1978). The existence of such feature detectors was thought to be indicated by Eimas and Corbit's (1973) adaptation experiments in which when listeners were asked to categorize members of a synthetic voice onset time continuum, it was demonstrated that the perceptual boundary between voiced and voiceless categories was shifted by repeated exposure to (that is adaptation with) either of the endpoint stimuli. The repeated stimulus exposure was thought to tire the neural feature detector. Highly specialized in-built feature detectors probably do not in fact exist as cross-language studies have shown that speakers of different languages may place phonetic boundaries at different points along the same acoustic continuum. However, studies on infants indicate that some acoustic dimensions may be naturally divided into categories (Streeter, 1976; Lasky, Syrdal-Lasky, and Klein, 1975). But there is ample evidence that the category boundaries may be modified or forgotten as a result of linguistic exposure (Zlatin and Koenigsknecht, 1975). Nevertheless the category boundaries used in adult speech perception are

very well learned and hard to change. Attesting to this is much information of an anecdotal nature from second-language learning. Children below a certain critical age (10-12 years old) learn a new language very well. They can correctly perceive unfamiliar phonemes and reproduce them. Adults, on the other hand have great difficulty with these tasks and generally never learn to speak a second language without an accent.

In reviewing categorical perception we saw that as an acoustic dimension is varied experimental subjects often divide the sound class which has the experimental dimension of interest as a partial cue in a categorical manner. We have not yet discussed however what happens when two acoustic dimensions are varied simultaneously. Sawsuch and Pisoni (1974) conducted an experiment in which subjects were asked to classify synthetic stop consonant-/a/ syllables that were varied in eleven steps from voiced labial to voiceless alveolar, in one case as /ba/ or /ta/ and in another case as /ba/, /da/, /pa/, or /ta/. On the basis of the results of these experiments Sawsuch and Pisoni were able to demonstrate that voicing and place cues do not act independently, i.e. cue combination is non-linear. Working with Sawsuch and Pisoni's data Oden (1978) proposed a 'fuzzy logical' model of phoneme identification which he suggests takes place in the following way: first each feature is evaluated and given a fuzzy rating, i.e. the place of articulation cue might be rated as 0.7 alveolar and 0.2 labial (For a fuller discussion of Fuzziness see next chapter.) Secondly, prototype matching takes place in which fuzzy combination rules are used to find fuzzy ratings for each sound as various possible phonemes. Finally perception classification takes place in which the sound is classified on the basis of the relative degree to which it matches the various alternative phoneme prototypes as specified by the matching function. Oden found that his model was in fair agreement with

experimental results. Oden and others (Oden, 1977; Massaro and Cohen, 1976) have also successfully tested fuzzy logical models on other pattern identification problems which the brain solves. This suggests that the brain makes fuzzy rather than binary decisions at various stages of speech processing. It should be noted that such a claim is not necessarily incompatible with categorical perception. Although various levels of processing may be fuzzy the final decision as to what the phoneme is might occur on a probabilistic basis or might occur more simply with a requirement that the feature combination rule for each sound is such that at most one possible sound might reach the threshold for recognition. The problem of information integration is one which has not received much attention in the literature but is one which is of major importance in the building of automatic recognition systems. For it does not matter how well information is extracted from the speech signal if that information is not interpreted and integrated appropriately. As was pointed out at the beginning of the discussion on perception, just hearing a language spoken does not guarantee that it communicates anything to us unless the sound patterns of the language are intelligible to us and unless we can analyze the sounds within our linguistic framework.

1.4.7 Prosody

Before leaving the subject of perception there is one aspect that is extremely important in interpreting the meaning of a spoken phrase. This is prosody, a term describing the patterns of stress and intonation of a language. Intonation is achieved through pitch variation and is used to indicate overall sentence structure e.g. there are special intonation patterns for interrogative questions, rhetorical questions, statements, exclamations and so on. Special intonations also convey information about

the speaker's emotions. Stress is achieved through pitch, timing, and loudness variations and is used to enhance meaning within a sentence. Timing variations are also used to indicate verbal parentheses (O'Malley, Kloker, and Dara-Abrams, 1973) The importance of prosodic effects as an aid to speech recognition have been pointed out by Lea, Medress, and Skinner (1974) and O'Kane (1977). Such effects are useful at several levels. At the acoustic-phonetic level, if a word is stressed we know that the vowels in that form are unlikely to be reduced. The greatest use of prosodic information can probably be made at the higher levels however. Stress patterns are of use in finding beginnings and ends of words and, as was pointed out above, in indicating which are the key words in a sentence. Intonation patterns will indicate what kind of sentence is being spoken.

1.5 REVIEW OF AUTOMATIC SPEECH RECOGNITION SYSTEMS WITH PARTICULAR EMPHASIS ON THE ACOUSTIC-PHONETIC COMPONENTS OF SUCH SYSTEMS

This review of Speech Recognition Systems is basically an overview of the trends in acoustic-phonetic analysis for speech recognition. Specific points of interest will be considered in greater detail in later chapters.

Automatic speech recognition studies divide historically into three (timewise) categories - pre-ARPA, the ARPA project, post-ARPA. In 1971 the Information Processing Technology Office of the Advanced Research Projects Agency (ARPA) of the United States Department of Defence initiated a five-year research and development programme on speech understanding. The objective of this programme was to build speech understanding systems that would accept connected speech from many cooperative speakers of a General American Dialect. Input was to be through a good quality microphone in a quiet room. Tuning of the system to new speakers was to be allowed. The

system was to be able to handle a vocabulary of at least 1000 words but was allowed to work in a restricted task domain with an artificial syntax appropriate to that domain. The system was to be able to run in a few times real time on a 100 MIPS computer, and was to achieve less than 10% semantic error. Five major systems were started and several supporting projects were also funded. After two years, funding on only the three most promising systems was continued. Of these systems only one, the Carnegie-Mellon Harpy system met the requirements laid down in the initial objectives. It accepted connected speech from five speakers speaking in a computer terminal room. Twenty training sentences were used for each new speaker. The task on which the system operated was document retrieval;the vocabulary was 1011 words (Medress et al, 1976).

Before the ARPA project speech recognition/understanding was mainly concerned with very limited vocabulary isolated word recognition systems such as the digit recognition systems of Davis, Biddulph, and Balashek (1952), Denes and Mathews (1960). Another field of endeavour was the design of recognition algorithms for specific classes of sounds. Examples of this are found in the work of Halle, Hughes, and Radley (1957) for stop consonants, and Forgie and Forgie (1959) for vowels. As studies such as these progressed with the aim of advancing automatic speech recognition the difficulties involved in the extraction of information from the speech waveform were realized and the concept of using 'higher-level' information such as semantic, syntactic, and pragmatic information as well as phonetic decoding of the input acoustic waveform was seen to be a useful approach. This approach is the one that was emphasized by the organizers of the ARPA project and it is interesting to note just how well the ARPA systems performed using relatively unsophisticated acoustic-phonetic analysis techniques.

In the Harpy system (Lowerre, 1976) speech is digitized at 10kHz and then fourteen linear prediction coefficients are computed for every 10msec of speech. These 10msec segments are then grouped together if they are sufficiently similar and a template matching procedure takes place. 98 templates are used and minimum residual error metric distance measures (Itakura, 1975) are computed to find out how well the test segment matches each of the templates. The recognition strategy is basically a verification technique. All possible paths through the finite state automaton used are stored in the machine. Working in a strictly left-to-right fashion, the system uses the results of the template matching to decide which path through the grammar is most likely. Although the top-scoring template matches the expected template only 60% of the time for the best-scoring path through the network, the system has enough higher-level information encoded into it to be able to handle this and the overall semantic accuracy of the system is 95%.

Another automatic speech understanding system was developed at Carnegie-Mellon University under the ARPA project. This is the Hearsay II system (Erman, Hayes-Roth, Lesser, and Reddy, 1980) in which preliminary analysis involves manner-of-articulation categorization of segments defined by amplitude and zero-crossing parameters. After this a word hypothesizer lists all words consistent with the hypothesized syllable structure. The greatest problem with this approach is coarticulation across word boundaries. Nevertheless a word score of 70% is achieved at this stage of the processing.

In contrast to the CMU Harpy and Hearsay systems, the Systems Development Corporation system (Bernstein, 1976) obtains a phonetic transcription from formant frequencies and other parameters that have been

extracted from the waveform. Actually several alternative labels are found for each phonetic segment and all the data are placed in a structure called the A-matrix. The utterance is processed in a left-to-right fashion with a list of all possible sentence-initial words being generated at first. After this the control mechanism retrieves an abstract phonemic representation from the lexicon for each lexical hypotheses and computes expected phonetic variants. The A-matrix is then examined for possible matches to these phonetic hypotheses. This system is capable of rejecting a large fraction of word hypotheses not in the sentence but only at the cost of rejecting the correct word about 10% of the time.

The Bolt, Beranek, and Newman system (Woods, 1976) also derives a phonetic transcription from formant frequencies and other useful parameters. Alternative possibilities are ordered in a segment lattice after a series of several passes through the data each pass refining the phonetic transcription. First-choice correct labels are obtained about 52% of the time and the correct label is within the top five label choices 83% of the time. Extensive phonological rules were incorporated into the system to aid in locating word boundaries etc. The system examines the segment lattice for segment combinations which provide good matches to words in the lexicon. These seed words are then used as the basis for left and right searches to build up partial sentence hypotheses.

A system for acoustic-phonetic analysis of continuous speech was also developed at Lincoln Laboratories, MIT under ARPA funding (Weinstein et al., 1975). Although this system was never incorporated into an overall Speech Understanding System it was recognized as giving the best performance in terms of phonetic output (Klatt, 1977). Preliminary segmentation and segment classification based on linear prediction spectral analysis and

fundamental frequency estimation placed phonemes in one of the classes: vowel-like sound, volume-dip within vowel-like sound, fricative-like sound, stop consonant. More detailed classification algorithms included detection and identification of some dipthongs, semi-vowels and nasals, a vowel identifier, a fricative identifier, and a stop consonant classifier based on analysis of bursts.

A system which in some ways is based on a similar philosophy to the Lincoln Laboratories system is the system developed at the University of Turin, Italy (De Mori, Rivoira, and Serra, 1975). This system which was commenced during the time of the ARPA project is still under active development and is characterized by very careful acoustic-phonetic classification. After segmentation, a pre-categorical classifier allocates phonemes to one of the classes sonorant or nonsonorant. The sonorant class is then subdivided into the classes vowel, semivowel, nasal, and liquid, while the nonsonorant class is divided into plosives and fricatives. Place of articulation and voicing classification comes next. At this stage contextual information is used to obtain very accurate results for sound classification. Examples of parts of the Turin system are found in Chapters 3 and 4.

Several systems which attempt to interpret the speech waveform in the same manner as does the human ear have been proposed in recent years. Alinat in France (1978) has shown how an artifical cochlea can be used in plosive consonant identification. Zwicker, Terhardt, and Paulus (1979) in Germany have proposed a system which relies on pre-processing in terms of auditory parameters. And Searle, Zachary Jacobson, and Rayment (1979) have developed a system for plosive recognition which is also auditorily based.

The major recent development in acoustic-phonetic models is the LAFS (Lexical Access From Spectra) proposed by Klatt (1979). Klatt claims that 'a bottom-up method of lexical access is an essential part of the normal speech decoding process'. He sees eight major problems for any such system:

(a) acoustic-phonetic invariance

(b) segmentation of the signal into phonetic units

(c) time normalization

(d) talker normalization

(e) lexical representation for optimal search

(f) phonological recoding of words in sentences

(g) dealing with errors in the initial phonetic representation during lexical matching

(h) interpretation of prosodic cues to lexical interpretation of sentence structure

LAFS is a system which, like Harpy, bypasses any phonetic coding stage by having a decoding network of spectral templates for all words in the lexicon. The templates inherently account for intraword coarticulation, and account for interword coarticulation by having a series of alternative possible spectra for word beginnings and terminations according to which sounds can occur at the beginnings and ends of words surrounding the word being investigated. Klatt claims that LAFS could provide a model of how human perceptual processes work and in particular the idea of bypassing the phonetic code would explain the experimental results of Savin and Bever (1970). He also acknowledges that there are cases where there is need for phonetic analysis (e.g. a new word not stored in the lexicon) when phonetic decoding is necessary. To solve this problem he proposes a phonetic

decoder called SCRIBER which could operate as an alternative path to LAFS when necessary. SCRIBER is interesting in that instead of using conventional phonemes as its basic segments for recognition it uses diphone units that extend from the middle of one phonetic segment to the middle of the phonetic segment following. Templates of all diphones are stored and recognition occurs by a residual maximum error distance estimation technique. Talker normalization is the same as in Harpy where twenty training sentences per new speaker are used to see if special speaker dependent templates are necessary to account for the idiosyncratic pronunciations of a particular speaker. The main advantage of SCRIBER is that the use of diphones means that explicit phonetic segmentation is not necessary and coarticulation effects between adjacent sounds are automatically accounted for.

1.6 THE FOPHO MODEL FOR AUTOMATIC SPEECH RECOGNITION AND ITS RELATION TO OTHER SPEECH RECOGNITION SYSTEMS

To motivate the need for this model, let us consider a phonetician who wishes to study a language which is not his native tongue and in which he is not fluent. In order to obtain a first-hand knowledge of this language, X, he decides to visit the country, Y, where X is spoken. So he does a crash course in X and obtains an idea of the grammatical structure of the language and learns an enough-to-just-get-by vocabulary and also a few set phrases. On his arrival in Y he finds that no-one speaks his native tongue and that he has to rely on his extremely limited knowledge of X in order to communicate. As a phonetician, he is trained to listen carefully. He tries to write down in phonetic notation what he has heard so that he can try to work out what is being said. Once he achieves a phonetic transcription he can then study it and try to discover where word

boundaries are. Then with the aid of a dictionary and grammar he can work out what was said. He will probably know that there were some mistakes in his transcription but he will have a fair idea of what was said. The phonetician differs from the casual foreign visitor in that, professionally, he wants to hear how the language is spoken and therefore listens carefully for any clues that will reveal the peculiar features of the language. The casual foreign visitor, on the other hand, will perceive the language in a manner more nearly resembling the manner in which he perceives his own language - which is to attend, not too carefully, to the actual sounds but to hear enough to be able, with the aid of other sources of information (e.g. grammar and pragmatics) to work out what is being said.

The phonetician knows about linguistic processes. He is aware of such effects as coarticulation, in which a sound is produced in a way which reflects the influence of neighbouring sounds. He will know about categorical perception, the effect whereby members of a class of sounds are always perceived as being definitely one member or another and never as part way between the two. The phonetician will realize that he has certain things to learn - effects that he has never previously encountered and essential features that characterize the language; that he must be willing to change hypotheses easily by designing his framework for the language to be sufficiently flexible. He will find that his transcription improves if he includes prosodic effects.

An automatic acoustic-phonetic recognizer of continuous speech must be very similar to our foreign phonetician. It must be able to perceive what is being said and produce a string of phonetic syllables. Although it should be able to perceive all the effects that the human brain can

perceive, it must pay more attention than the human brain usually does to every sound that is produced. Also it must formally know something of the structure of the language that it is attempting to recognize. This is information that is hardwired into the system. It must be more like the phonetician than the casual listener because it must be able to go from the speech waveform (which both hear) to a consciously formal representation of what is heard. And where there is ambiguity and uncertainty in what is heard it must have enough knowledge of phonetic structure to resolve it.

The Foreign Phonetician model (FOPHO) for automatic speech recognition differs from the recognition systems discussed in the previous section in several ways. While the development of this model is an attempt to model human adult speech perception it is the perception of the foreigner rather than the native that is being modelled. No Speech Understanding System is going to be able to incorporate (given the current state of the art) all the linguistic information of a language that a native speaker has at his disposal by the time he is mature.

At the acoustic-phonetic level the system proposed here while addressing itself to the same problems as Klatt's (1979) system, will be significantly different in structure. Like the Turin and Lincoln Laboratories systems the FOPHO acoustic-phonetic decoder is hierarchic i.e. the speech waveform is segmented into broad phonetic classes with finer and finer classifications taking place at secondary stages. At all levels the classification rules take coarticulation into account. A hierarchic system allows for systematic backtracking if a higher stage level of processing indicates that the phonetic result is wrong or at least unlikely. Also as the Ali et al (1970) experiments illustrate, when a sound is excised or partially masked, some information about it can still be gleaned from

coarticulation effects in surrounding phonemes. Thus a sound with a 'bad' classification might have been misclassified at one of many stages. It might have been misclassified according to voicing, manner, or place of articulation. By examing results at these levels we can select the most doubtful result and look at possible alternatives.

As a matter of fact the integration of information about the various features will take place in a manner somewhat similar to that proposed by Oden (1978) for the brain. Differences with Oden's model are that FOPHO would attempt to achieve categorical-like feature recognition and all the recognition stages would be fuzzy. (How this can be done is demonstrated in Chapter 2 where the principle for fuzzy recognition is introduced.) Thus in a consonant recognition procedure only one phoneme along with its fuzzy rating will be passed to the word recognition system for initial consideration (unless the word recognition processor requests a list of alternatives as it well might in a word initial situation). However details of the procedure are retained in memory for a time so that inconsistencies occurring 'higher-up' can be checked. Vowel recognition is somewhat different from consonant recognition - here the feature results are not so categorically defined and the possibility of the unknown sound being several vowels will be entertained even at the word recognition stage. (Details of this are to be found in Chapter 5.)

At the word recognition stage FOPHO would follow the word-recognition model of Marslen-Wilson and Welsh (1978). A phonological rule controller would act on the dynamically incoming phonetic transcript with which there will also be some information about word boundaries from juncture rules (see Chapter 6) and also from the concomitant output from the prosodic decoder. The phonological controller then guides interaction with the

lexicon. It can also request information from the acoustic-phonetic decoder. In the initial stages of word identification the phonological controller uses the information from the phonetic transcript to alert possible words in the lexicon. As more and more of these possibilities are eliminated and the phonological controller is reasonably certain of the identity of the word being tested it will tell the acoustic-phonetic decoder to temporarily stop identifying new incoming phonemes and to start checking to see if the next few phonemes are those hypothesized by the phonological controller. If the match is reasonable the acoustic-phonetic controller will be directed to resume standard identification and decoding of the input for the next word; if it is not the acoustic-phonetic decoder will be asked to supply an alternative hypothesis.

A question might be asked here - if the input is truly sequential, how does the acoustic-phonetic decoder take backwards coarticulation into account? The fine decoding of a sound (i.e. voicing and place) information is probably not undertaken until the beginning of the following sound is input (Even the times for the fastest of Marslen-Wilson's shadowers (1975) would allow for this) so what is happening is that the decoder is probably working in parallel on several sounds at different levels. Thus when necessary a final result on one sound might await further knowledge about the following sound so that coarticulation effects can be used in the final decision.

Where the system displays its foreigness is that often because inadequate knowledge of the language is built into the system it will make many more mistakes than a human being hearing his native language would. Thus there will have to be a lot of backtracking and rechecking and general to-ing and fro-ing between the acoustic-phonetic decoder and the

phonological controller. However as the causes of the system's slowness and inaccuracies are examined, new rules can be added and old ones modified so that the system will not have to use its backtracking facility so much. And when the system finds itself hopelessly lost it would be able to ask for a repetition or clarification from the speaker.

The remainder of the thesis concentrates on aspects and design of the acoustic-phonetic decoder. This chapter has aimed at setting the acoustic-phonetic decoding process in a wider framework and suggesting a new model, FOPHO, for automatic speech recognition. Admittedly only the understanding process up to word recognition has been discussed but it should be obvious that the concept of understanding by a foreigner (particularly a foreigner who knows what he is listening for such as a linguist) can be profitably applied at phrase recognition level as well.

Chapter 2

ON SYNTACTIC PATTERN RECOGNITION AND FUZZY SET THEORY

2.1 SPEECH RECOGNITION AS A PROBLEM IN SYNTACTIC PATTERN RECOGNITION WHETHER PERFORMED AUTOMATICALLY OR BY THE HUMAN BRAIN

Speech recognition and understanding is a problem of pattern recognition. There are two basic methods of tackling any pattern recognition problem. In the approach known variously as 'statistical' or 'decision-theoretic', each pattern is represented by a feature vector and recognition of a pattern is made by partitioning feature space. In the syntactic approach each pattern is represented as a composition of its component sub-patterns, called primitives. In this case, recognition of a pattern is usually made by parsing the pattern structure according to a given set of syntax rules. The syntactic approach is the one that is predominantly used in this thesis.

Figure 2.1 is a diagram of the process of implementing and operating a syntactic pattern recognition system. In this thesis the main thrust of the work is at the analysis stage (above the dotted line in Figure 2.1) of selecting suitable primitives and developing structural rules. This analysis work has mainly been performed manually and the results of the analysis are used to develop two types of automatic recognition rules, one set for automatic extraction of suitable primitives and the other for the automatic analysis of the pattern according to the structural rules.

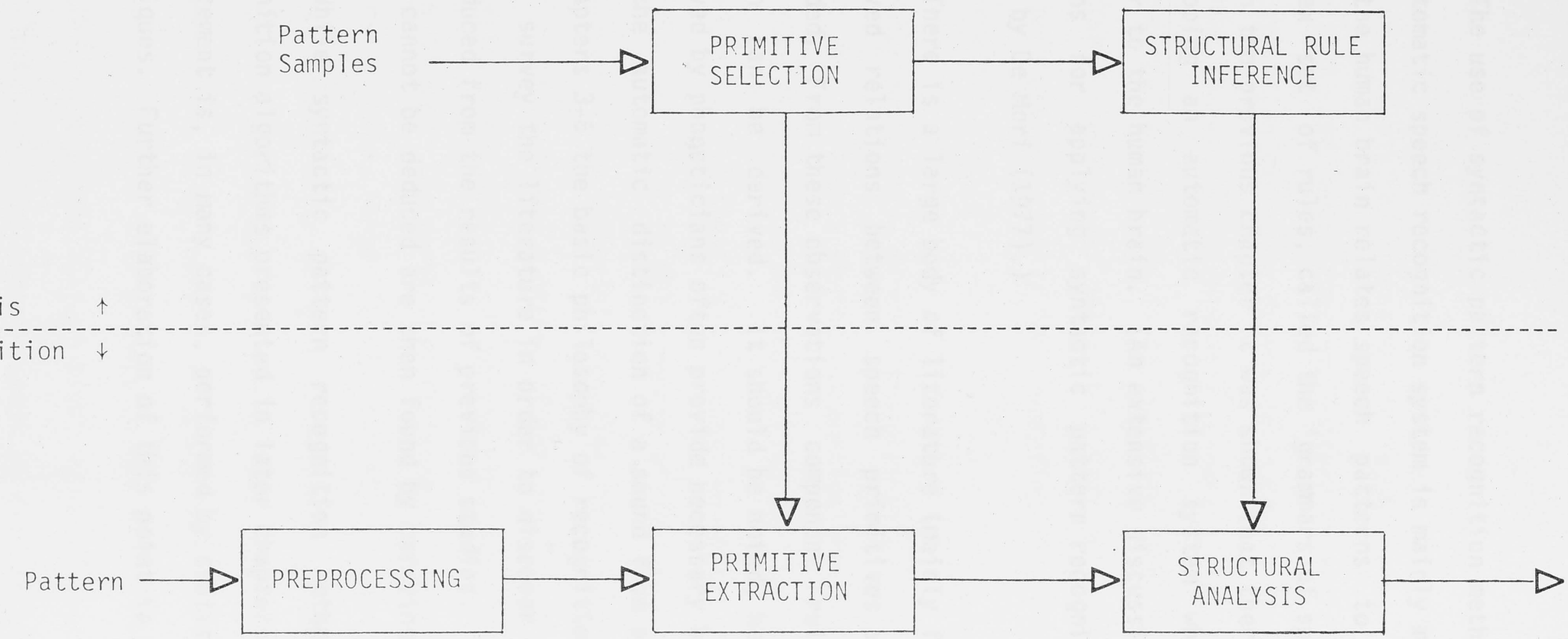


FIGURE 2.1: Structural pattern recognition

The use of syntactic pattern recognition methods in the development of an automatic speech recognition system is mainly motivated by the knowledge that the human brain relates speech patterns to linguistic items by a complex set of rules, called the 'grammars of speech' by Liberman (1970). And in the previous chapter it was shown that there are good reasons for developing an automatic recognition system which operates in a similar manner to the human brain. (An extensive discussion of the motivational reasons for applying syntactic pattern recognition methods to speech is given by De Mori (1977).)

There is a large body of literature (mainly from Phonetics) in which observed relations between speech primitives and linguistic entities is recorded. From these observations component rules of the grammars of speech can be derived. It should be noted, however, that the relations observed by phoneticians often provide necessary but not sufficient rules for the automatic distinction of a sound from all other types of sounds. In Chapters 3-6 the basic philosophy of recognition rule design has been to first survey the literature in order to discover to what extent a rule can be deduced from the results of previous studies. Those parts of the rule which cannot be deduced are then found by carrying out special experiments.

While syntactic pattern recognition methods are used in the recognition algorithms presented in later chapters, primitive selection and measurement is, in many cases, performed by statistical pattern recognition techniques. Further elaboration of this point is given in Section 2.3.

2.2 FUZZY SET THEORY

Human decision processes in such tasks as continuous speech recognition and understanding, pattern recognition, and conceptualization and abstraction typically depend on vague and imprecise data. Yet these decision processes are eminently successful in allowing humans to interact with the real world. And in tasks where the data available is vague, imprecise, complex and noisy it is very noticeable that humans always far excel machines and usually feel justifiably confident that their decisions are correct.

Fuzzy set theory, first introduced by L. A. Zadeh in 1965, is an attempt to provide a strict mathematical framework within which imprecise conceptual phenomena in decision making may be precisely and rigorously studied. In classical set theory an object either IS a member of some particular set or it IS NOT a member of that set. Using Boolean values, the value of the membership function for some item in a particular set is either 0 (if it is not a member of that set) or 1 (if it is a member of that set). An object can, on the other hand, belong to a fuzzy set to a certain degree. The membership function of a fuzzy set can take values anywhere in the range [0,1]. For example, in Figure 2.2 the membership function for the fuzzy set of tall people as a function of height is given. Note that a person who is 2 metres tall belongs to the set of tall people with membership 1, as there is no doubt that that person is tall. A person who is 1.7 metres tall belongs to the set of tall people with membership 0.3. This is a person that we might commonly describe as 'not terribly tall', that is, this person cannot be classed as being tall with the same certainty that the person who is 2 metres tall can.

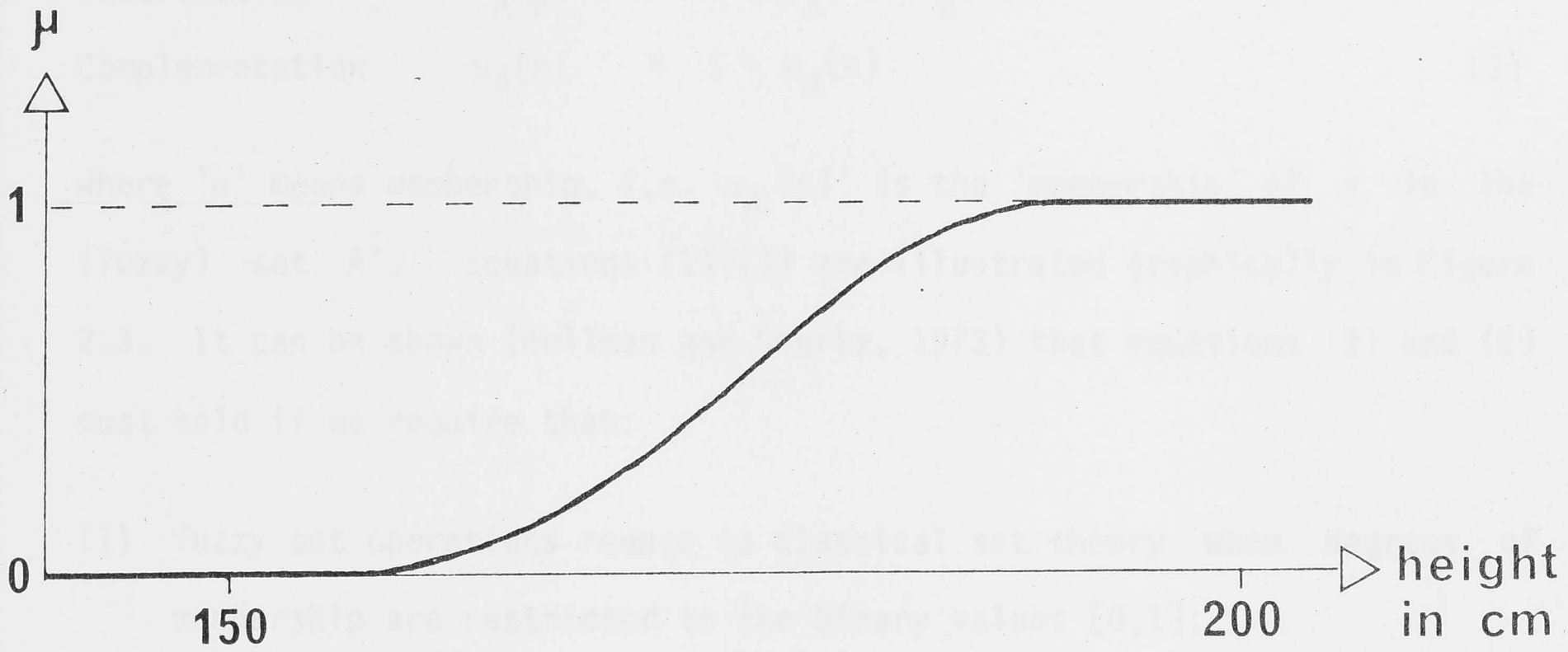


FIGURE 2.2: Membership function of the fuzzy set of tall people

Membership functions resulting from operations on fuzzy sets are generally defined as follows:

Union	$\mu_{A\cup B}(x)$	=	$\max(\mu_A(x), \mu_B(x))$	(1)
Intersection	$\mu_{A\cap B}(x)$	=	$\min(\mu_A(x), \mu_B(x))$	(2)
Complementation	$\mu_{\overline{A}}(x)$	=	$1 - \mu_A(x)$	(3)

where 'μ' means membership, i.e. $\mu_A(x)$' is the 'membership of x in the (fuzzy) set A'. Equations (1)-(3) are illustrated graphically in Figure 2.3. It can be shown (Bellman and Giertz, 1973) that equations (1) and (2) must hold if we require that:

(1) fuzzy set operations reduce to classical set theory when degrees of membership are restricted to the binary values [0,1];

(2) the natural numerical order relation of degrees of membership is to be consistent with one's concept of union and intersection;

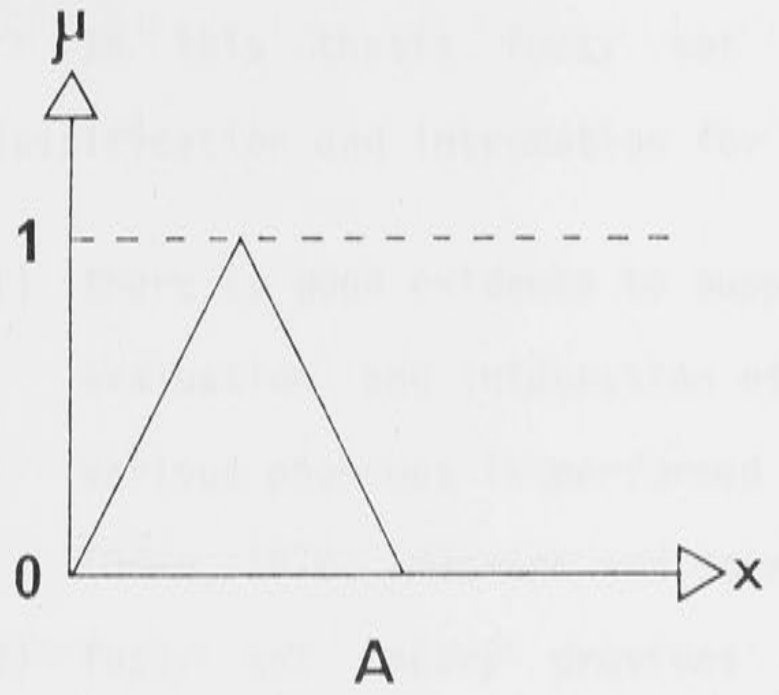
(3) $A\cup B$, $A\cap B$ are continuous, nondecreasing in A, B;

(4) associativity, commutativity, idempotency and distributivity hold on the union and intersection operations.

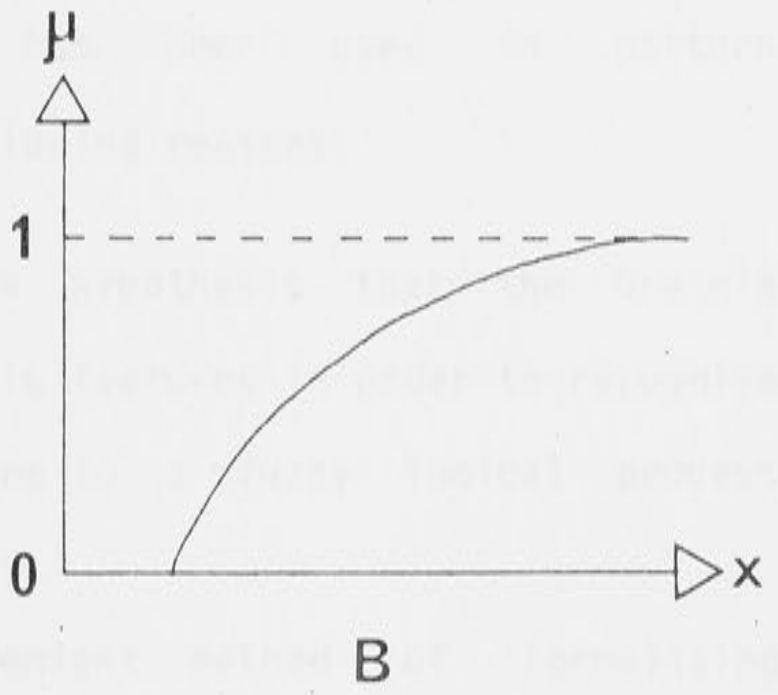
Complementation is not so constrained and there are occasions when the definition given in (3) is somewhat unsatisfactory (Gaines, 1976), although for the applications described in later chapters it has been found to be adequate. It should be noted that, unlike in classical set theory:

$A \cup \overline{A} \neq X$, where X represents all members of the universe under consideration

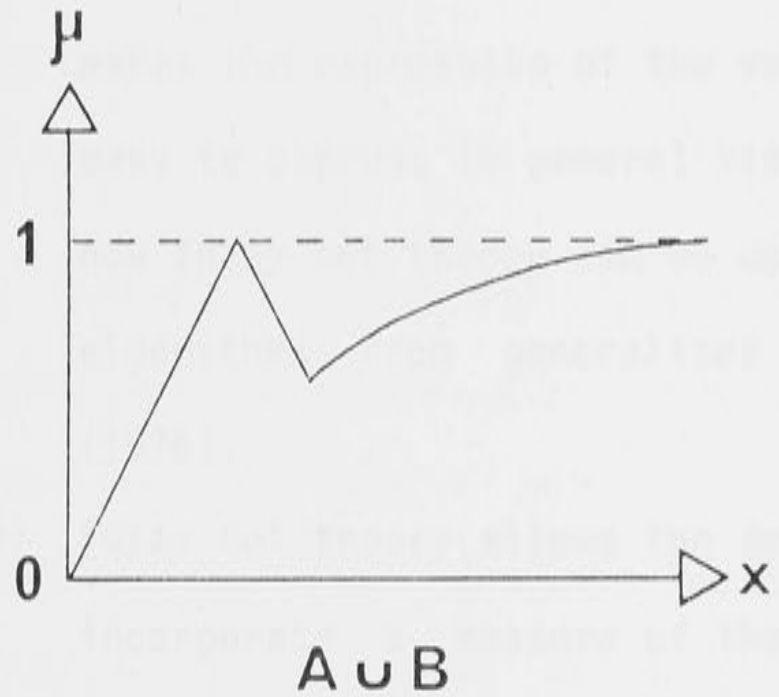
$A \cap A \neq 0$



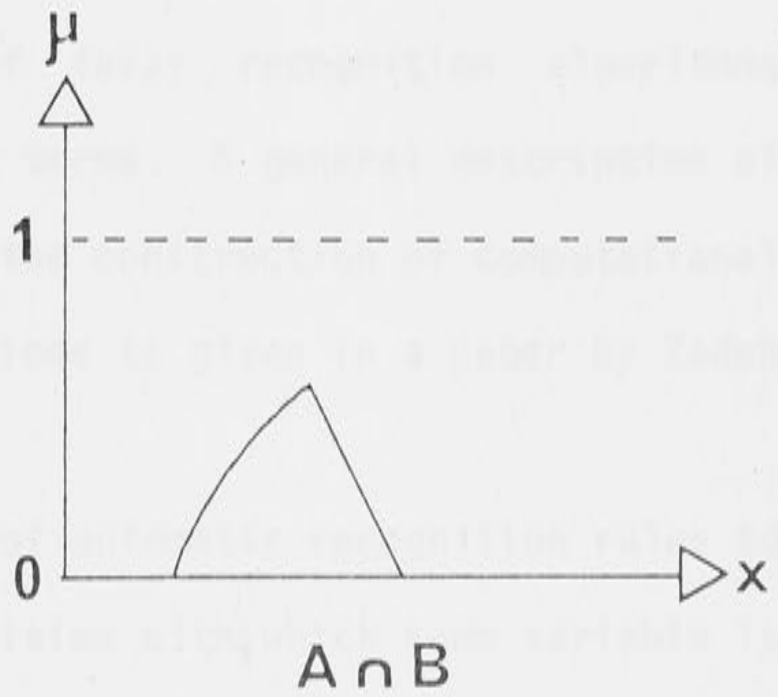
A



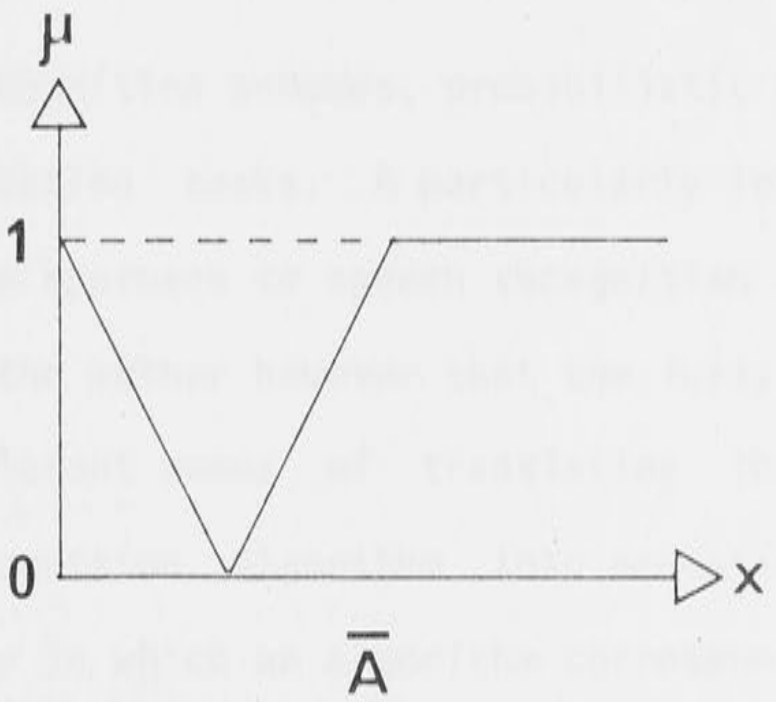
B



A ∪ B



A ∩ B



$\overline{A}$

FIGURE 2.3: Examples of the operations union, intersection, and complementation on fuzzy sets.

In this thesis fuzzy set theory has been used in pattern classification and integration for the following reasons:

(1) there is good evidence to support the hypothesis that the brain's evaluation and integration of acoustic features in order to recognise various phonemes is performed according to a fuzzy logical process (Oden, 1978; Massaro and Oden, 1980);

(2) fuzzy set theory provides a convenient method of formalising acoustic-phonetic rules for computational processing. Conversely, it makes the expression of the results of fuzzy recognition algorithms easy to express in general linguistic terms. A general description of how fuzzy set theory can be used for the construction of computational algorithms from generalised assertions is given in a paper by Zadeh (1976).

(3) fuzzy set theory allows the designer of automatic recognition rules to incorporate a measure of the imprecision with which some variable is being measured into the recognition scheme.

In many speech recognition schemes, probabilistic measures are used in the pattern classification tasks. A particularly interesting example of probabilistic inference approach to speech recognition is given by Jelinek (1976). It seems to the author however that the fuzzy set theory approach provides a much more elegant means of translating the operation of an automatic speech recognition algorithm into acoustic-phonetic terms and thus monitoring the way in which an algorithm corresponds at its various stages to recognition methods used by the human brain.

The discussion of fuzzy set theory presented here has, necessarily, been extremely brief and incomplete. An extensive bibliography of works detailing the theory and applications of fuzzy sets has been compiled by Gains and Kohout (1977).

2.3 OVERVIEW OF THE FUZZY SYNTACTIC RULE APPROACH TO CONTINUOUS SPEECH RECOGNITION ADOPTED IN THIS THESIS

The interpretation of speech patterns involves the generation of hypotheses concerning possible phonemic transcription of syllable segments automatically extracted from a numerical representation of energy-frequency-time data obtained by short-term spectral analysis of a spoken sentence. Each hypothesis is evaluated and a degree of trustworthiness is assigned to it in such a way that it can be further processed for generating and coherently evaluating hypotheses about the words, the syntactic structure, and the semantics of the spoken sentence.

Methods based on fuzzy restrictions are proposed in later chapters for extracting acoustic features (such as vowel formant measurements) from the description of acoustic patterns. The mechanics of doing this are discussed in the following section. Methods based on fuzzy relations are also proposed for relating acoustic features to phonetic and phonemic interpretations. As well, the use of restrictions and relations to compute the possibility of a hypothesis (a phonetic or phonemic interpretation of a speech pattern) is illustrated. The approach proposed here and the rules described in the following chapters are an attempt to formalise the intuitive logic used by a foreign phonetician.

Evaluating the possibility that any particular phonetic hypothesis is associated with $p\{t_i,t_j\}$, the acoustic waveform associated with the time interval (t_i,t_j), involves consideration of a composite question. From Zadeh's theory of the definition of complex imprecise concepts (see Zadeh, 1976 for details) a composite question defining a complex imprecise concept

is a triple

$$Q \triangleq (U,B,A)$$

where U is the universe of discourse (in our case, the set of all possible acoustic patterns), B is a structured linguistic variable, and A is a set of possible answers.

The structure of the linguistic variable is represented by a set of fuzzy rules relating a phoneme or phonetic hypothesis to acoustic features or other previously generated hypotheses.

A fuzzy rule (see Zadeh, 1976 again, for a detailed introduction of this item) is a syntactic rule with a 'degree of grammaticality' attached to it. Let S_T be the set of such rules. Some of the acoustic features may be considered as belonging to a terminal alphabet V_T. The other features are related to those in V_T by rules and belong to the non-terminal alphabet V_N.

Each element in V_T is a fuzzy linguistic variable, i.e. a triple (Y,U,R(Y)) where Y is a label for a fuzzy set representing a restriction R(Y) on the patterns p belonging to the universe U. Examples of such fuzzy sets and restrictions will be introduced in the following sections.

Fuzzy restrictions define the MEANING of linguistic variables in U.

When fuzzy rules are applied, relations are established between elements in V_N and strings in V_T.

In our approach, some elements of V_N are phonetic or phonemic hypotheses and we are interested in their degree of compatibility between them and a pattern $p\{t_i,t_j\} \in U$.

A distribution of degrees of compatibility between an element H V and the patterns of U is a fuzzy restriction.

Usually this distribution is not known *a priori* and it is not convenient to compute a priori its value for the infinite elements of U. Rather, when a particular pattern is analysed, its degree of compatibility with the hypothesis H can be calculated by applying so-called 'semantic rules' associated with the syntactic rules. Let S_N be the set of such semantic rules. The rules of S_N allow one to evaluate the meaning of a complex concept built by the application of rules in S_T. Also the rules in S_N may be used for obtaining evidence of a complex hypothesis $H \in V_N$ given evidence of the features described by variables in V_T which contributed to the generation of H.

2.4 PRIMITIVE SELECTION

The choice of primitives for speech pattern classification and phonetic hypothesis evaluation is dictated largely by those parameters which have been found to be effective for human perception of speech and by those features of speech patterns which are easy to extract and measure accurately by automatic means.

2.4.1 Preprocessing

Many of the primitives for speech pattern classification are parameters of the frequency transform of the digitized speech waveform. From this frequency transform it is possible to measure formants (concentrations of energy in the frequency domain corresponding to the resonances of the vocal tract) and the spectra of plosive bursts and fricative noise.

To obtain a frequency transform of the digitized speech waveform two techniques are commonly used. One is the Fast Fourier Transform (Cooley and Tukey, 1965) and the other is Linear Predictive Coding (Atal and Hanauer, 1971). Linear Predictive Coding inherently provides a vocal tract model and is probably the more popular method for speech processing as it provides an algebraic method of finding the formants of many sounds and also produces smooth spectra. Nevertheless, a combination of the two methods gives the most satisfactory results, particularly when one is dealing with sounds in whose production the nasal tract is important.

The basic processing for the experiments described in Chapters 3 and 4 was carried out at the University of Turin. In the processing system there, the results from Fast Fourier Transform and Linear Predictive Coding analyses are compared to obtain values of formants. Other parameters extracted include some waveform timing parameters and parameters giving the energy of the waveform in different frequency bands. A recent and very useful addition to this system has been an algorithm for automatic formant tracking (Laface, 1980).

The processing for the work described in Chapters 5 and 6 was carried out partially with a suite of Linear Prediction programs running on a DEC-10 computer (O'Kane, 1976) and partially using a commercially available software package known as the Interactive Laboratory System (ILS), produced and marketed by Signal Technology Inc. (Pfeiffer, 1978). The main processing method used by ILS is Linear Predictive Coding, although Fast Fourier Transform programs are also provided. Using ILS it is also possible to measure parameters relating to energy in various frequency bands, and timing.

2.4.2 Choice and Classification of Primitives

In the following four chapters there is considerable discussion on what parameters constitute suitable primitives for the classification of various classes of sounds. Because of the hierarchical structure into which the algorithms discussed in this thesis are intended to fit (see Section 1.6), the pattern primitives at various stages of recognition are sometimes different. Thus at the early stages of segmentation the pattern primitives are mainly parameters concerning the energy in certain frequency bands, while at the higher levels of classification they are often formant-type measurements.

Once the appropriate primitive has been chosen, the methods of classification used in this thesis generally take one of two forms. In the first form the primitive falls within a certain region in one dimension. It is classified as belonging (with varying degrees of membership) to the fuzzy sets defined over that dimension i.e. the fuzzy restriction on that dimension. Figure 2.4 illustrates the manner in which a series of overlapping fuzzy sets are defined over the dimension x which is the dimension in which a value of a certain primitive is being sought. The point z on the x-axis thus would belong to the different fuzzy sets as follows:

μ (z) = 0.6

μ (z) = 1.0

μ (z) = 0.4

μ (z) = μ (z) = 0.0

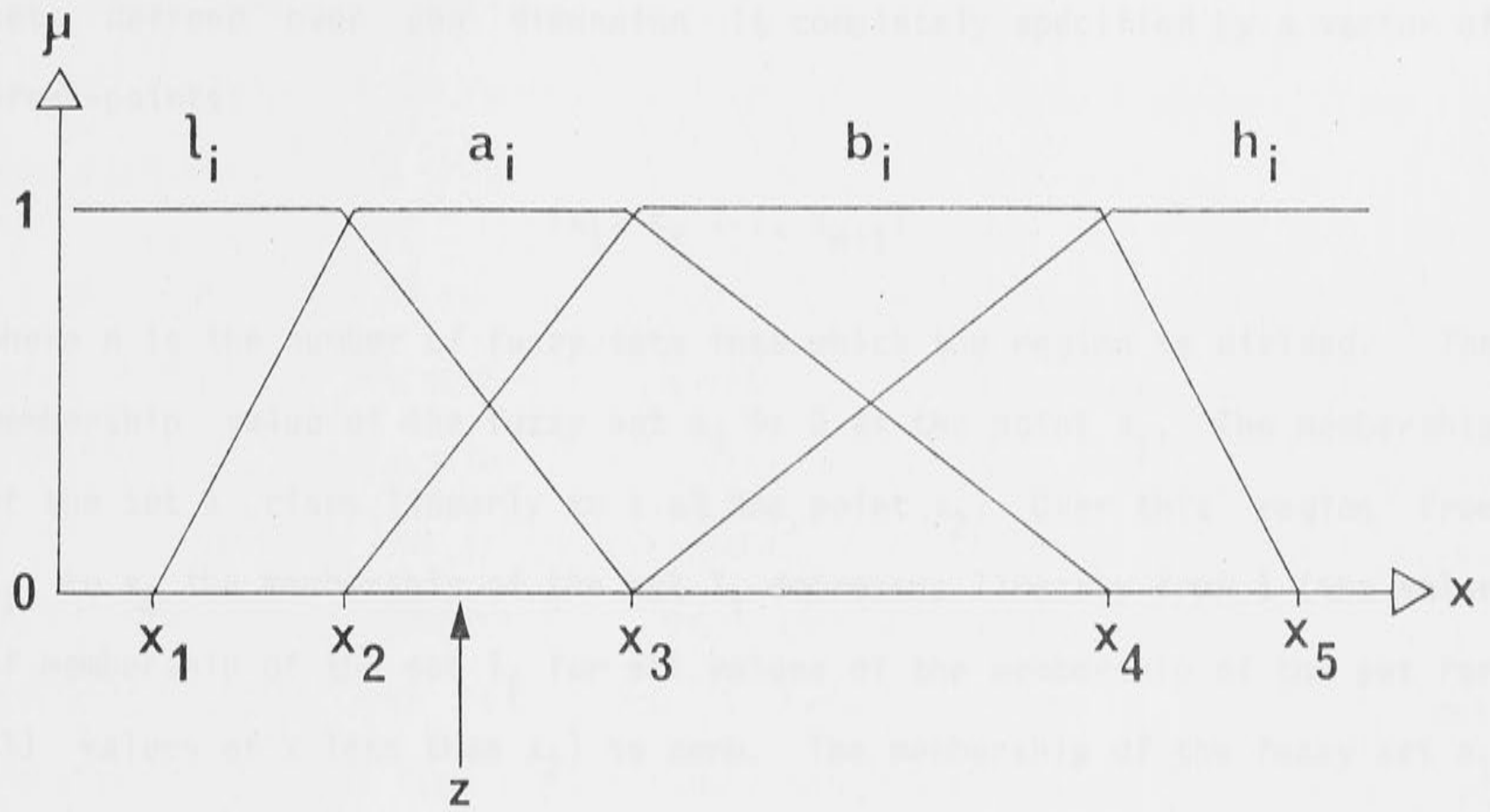


FIGURE 2.4: Membership functions of overlapping fuzzy sets defined over the dimension x.

The fuzzy sets defined in one dimension are named alphabetically from left to right except for the lowest and the highest sets which are called l_i and h_i respectively (i is a dummy subscript). Each set of these fuzzy sets defined over one dimension is completely specified by a vector of break-points:

$$\{x_1, x_2 \ldots, x_{n+1}\}$$

where n is the number of fuzzy sets into which the region is divided. The membership value of the fuzzy set a_i is 0 at the point x_1. The membership of the set a rises linearly to 1 at the point x_2. Over this region from x_1 to x_2 the membership of the set l_i decreases linearly from 1 (the value of membership of the set l_i for all values of the membership of the set for all values of x less than x_2) to zero. The membership of the fuzzy set a_i remains at the value 1 for the region from x_2 to x_3 from whence it decreases linearly to 0 over the region from x_3 to x_4. The membership of the fuzzy set b_i is zero for all values of x less than x_2, rises linearly to 1 over x_2 to x_3, remains at 1 from x_3 to x_4, and decreases linearly to 0 over the x_4 to x_5 region. And so on.

It should be noted that all the foregoing merely describes a convenient way of partitioning some particular region for classification purposes. It is by no means a unique method of partitioning a region, nor it is always the most suitable method of defining the partition. However, it is used for one-dimensional partitions in this thesis to provide a consistent method of rule presentation. There are examples of this case in Chapters 3, 4, 5 and 6.

The second form is the two-dimensional analogue of the case described above. It is used when two-dimensional feature space can be conveniently partitioned into rectangular partitions. Feature space is then defined by two one-dimensional partitions along the x and y axes. The fuzzy AND rule is used to describe the rectangular fuzzy sets in feature space. There are many examples of this in Chapters 3, 4 and 6.

A modification of the second form is used when two-dimensional feature space cannot be conveniently partitioned into rectangular regions. It might be used, for example, if one wished to define a circular, elliptical, or odd-shaped fuzzy set in feature space. For such a set there is a region within which fuzzy membership is one. Surrounding this region is another region in which the membership is between 0 and 1, or in which the membership decreases from 1 to 0 depending on the distance from the region of high membership. An example of this is given in Figure 5.10.

2.5 FUZZY AUTOMATA

A fuzzy automaton provides a description of the way in which a pattern described in terms of fuzzy primitives is (automatically) classified as being a particular pattern. The finite fuzzy automata used in this thesis are completely specified by the following parameters:

I: a finite set (classical) of input states

Q: a finite set of internal states

V: a finite set of output states

f: a function defining the fuzzy transition between any two states for some specified input value

A special subroutine (AUTMN) is used for all the automata computed in this thesis. The parameters listed above are passed to the subroutine along with the pattern to be classified, and the memberships of the output states are returned.

Chapter 3

DEVELOPMENT OF A CONTEXT DEPENDENT ALGORITHM TO DISTINGUISH BETWEEN LIQUID AND NASAL CONSONANTS

3.1 DESCRIPTION OF THE LIQUID/NASAL ALGORITHM DEVELOPED BY DE MORI, LAFACE, AND TORASSO, 1977

A suggested component of a Speech Understanding System which separated two subsets of the sonorant consonants, the liquids /l/ and /r/ and the nasals /n/ and /m/, has been described by De Mori, Laface and Torasso (1977). This component occurs in the SUS hierarchy of De Mori, Laface and Piccolo (1976) after a precategorical classification in which the continuous speech is segmented according to a syntactic grammar into pseudo-syllabic nuclei and hypotheses are proposed for each sound. According to these hypotheses the sound is placed in one of the categories vowel, sonorant consonant, or non-sonorant consonant.

The rules were devised from examination of speech parameters for all possible VCV (vowel-consonant-vowel) coarticulations in Italian, under the restriction that the consonant was either a liquid or a nasal. After each VCV pseudo-syllable had been extracted from continuous speech the formants were carefully tracked by a semi-automatic procedure, the energies associated with the first three formants were recorded, and some wide band energy measurements were made.

The rules for labelling a consonant as either liquid or nasal were syntactic rules, the scoring for which was designed according to fuzzy set theory . The overall objective of the rules was to classify successfully

each liquid or nasal sound in its appropriate category with membership greater than 0.5 and to ensure that no sound was classified in the incorrect category with membership greater than 0.5. With this philosophy the rules developed by De Mori, Laface and Torasso (1977) from data for one male speaker were:

$$\begin{aligned}\text{Liquid} = {} & l_1 + a_1([u,*]l_3 + [e,*](l_3 + a_3h_7) \\ & + [i,*](a_8 + h_8(l_4 + a_4h_7) \\ & + [o,*](l_5 + a_5h_7) \\ & + [a,*](l_6 + a_6(l_9 + a_9h_7)))\end{aligned}$$

$$\begin{aligned}\text{Nasal} = {} & h_1 + a_1([e,*](h_3 + a_3l_7) \\ & + [i,*]h_8(h_4 + a_4l_7) \\ & + [o,*](h_5 + a_5l_7) \\ & + [a,*]h_9(h_6 + a_6l_7)\end{aligned}$$

where the contexts of the square brackets give the contextual conditions and * indicates a 'don't care' condition, e.g. in the case of [a ,*] the context is /a/ before the consonant and any of the five Italian vowels after the consonant. The symbols $\{l_i, a_i, h_i\}$ (where i is a dummy subscript), refer to the fuzzy restrictions for various parameters which are represented by the symbols D12, D23, Dur(dip(S)), and Dur(dip(R)), in V_N, the non-terminal alphabet for acoustic features. These symbols and the fuzzy restrictions which operate on them are defined in Table 3.1. The vectors of break points associated with the fuzzy restrictions are also given in Table 3.1.

Parameter	Dimension	Fuzzy Restrictions	Associated Vector of Break Points
D12	dB	$\{l_1, a_1, h_1\}$	{3, 4, 15, 16}
D12	dB	$\{l_2, a_2, h_2\}$	{13.5, 14, 15, 15.5}
D12	dB	$\{l_3, a_3, h_3\}$	{9.5, 10, 12, 12.5}
D12	dB	$\{l_4, a_4, h_4\}$	{5.5, 6, 7, 7.5}
D12	dB	$\{l_5, a_5, h_5\}$	{8, 9, 12, 13}
D12	dB	$\{l_6, a_6, h_6\}$	{4, 5, 10, 11}
D23	dB	$\{l_7, a_7, h_7\}$	{-7, -6, 1, 2}
Dur(dip(S))	msec	$\{l_8, a_8, h_8\}$	{22, 20, 0, 0}
Dur(dip(R))	msec	$\{l_9, a_9, h_9\}$	{22, 20, 0, 0}

TABLE 3.1: Details of fuzzy restrictions on the parameters, D12, D23, Dur(dip(S)) and Dur(dip(R)). D12 is the difference in energy between the energy associated with the first and second formants measured at the point when the energy associated with the second formant has an absolute minimum in the syllable. D23 is the corresponding result for second and third formants. Duration(dip(S)) is the time interval for which the values of the short-time spectrum S fall below some specified value. Duration(dip(R)) is the corresponding result for R which is the ratio of the energy in the short-time spectrum in the 200-900 Hz band to the energy in the short-time spectrum in the 5-10 kHz band.

It should be noted that the rules just described have been presented in slightly different format from the way in which they were presented in the paper by De Mori, Laface and Torasso (1977). This was done in order to maintain a consistent method of rule presentation throughout this thesis.

3.2 COMPUTER IMPLEMENTATION

The author wrote a computer program to implement these rules. When the program was tested with data for several speakers it was found that the recognition rule structure needed considerable modification. The process of developing and testing the program to implement liquid/nasal consonant distinction is described in the remainder of this chapter.

For each VCV case considered, parameters were extracted and the parameter values and the rule strategy were passed to the general subroutine, AUTMN, for processing finite fuzzy automata.

The data from which the rules for the two main parameters were devised are displayed in Figures 3.1(a)-(e). Each figure shows D12 versus D23 for all cases in which the first vowel of the VCV combination was the same. Cases for which there was some doubt over the accuracy of the parameter measurements were excluded. 82% of the as nasals were recognized as nasals with membership 1.0 by the context-free condition h_1. However this condition incorrectly categorized 14% of the liquids as nasals with membership 1.0 although 4% of this same class of liquids were also recognized as liquids by context dependent rules. Such a case, where a sound is categorized as being in two mutually exclusive classes with membership greater than 0.5, is an indication that at least one of the rules is faulty. 20% of liquids were correctly categorized and no nasals were incorrectly categorized by the l_1 rule. The other recognition figures are set out in Table 3.2.

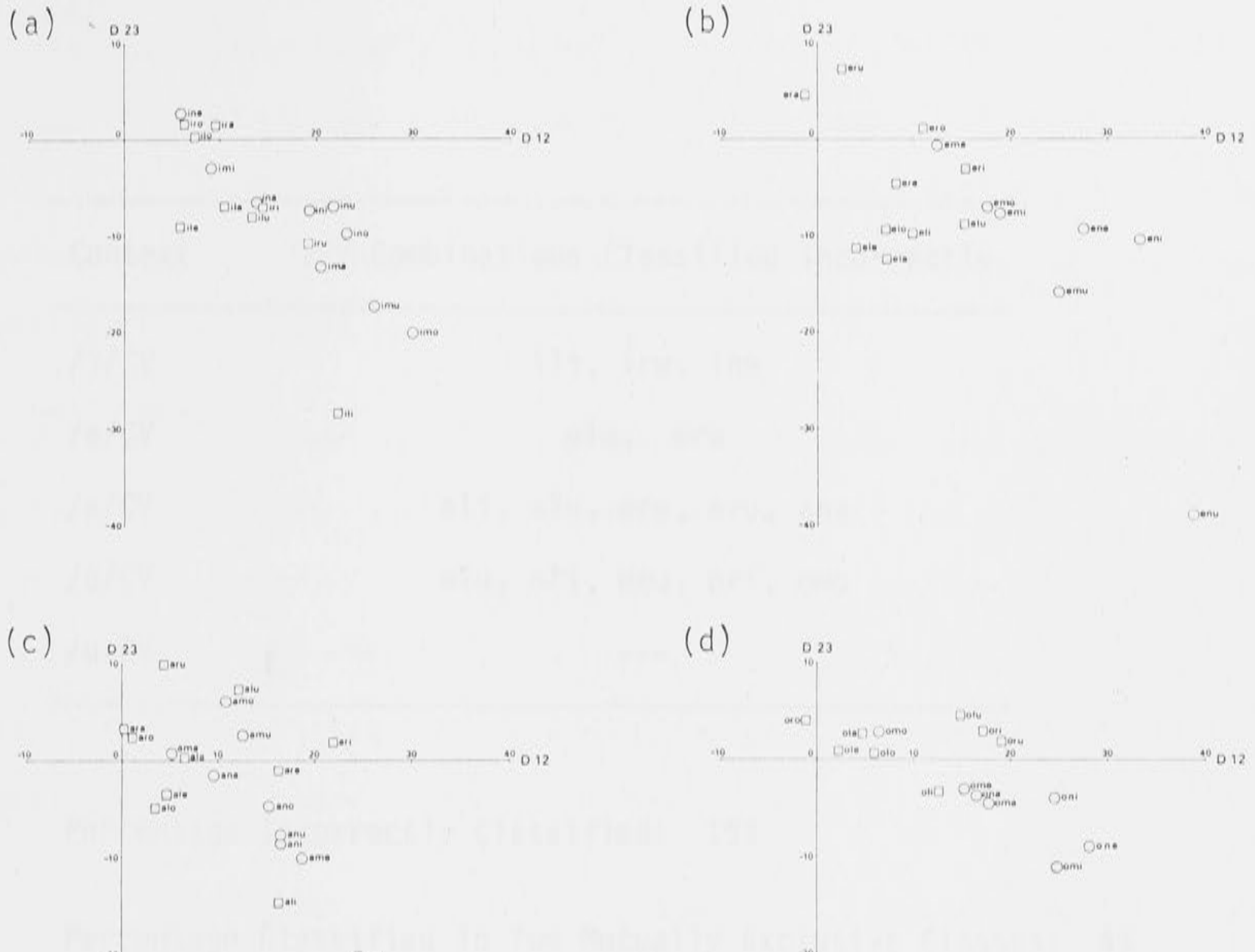


FIGURE 3.1(a)-(e): D23 versus D12 for the cases X-consonant-varying vowel, where X stands for the vowel /i/, /e/, /a/, /o/, or /u/ in Figures 3.1(a), (b), (c), (d) and (e) respectively. Data for one speaker.

Note: o is always a case of vowel-nasal consonant-vowel and □ is always a case of vowel-liquid consonant-vowel in all the figures in this chapter.

Context	VCV Combinations Classified Incorrectly
/i/CV	ili, iru, ine
/e/CV	elu, eru
/a/CV	ali, alu, are, aru, ana
/o/CV	olu, oli, oru, ori, omo
/u/CV	---

Percentage Incorrectly Classified: 15%

Percentage Classified in Two Mutually Exclusive Classes: 4%

TABLE 3.2: Recognition results for first version of liquid/nasal rules.

The recognition score was improved significantly by slight alteration of some of the numbers in the vector of break points. In particular, by making the h points of several of the questions somewhat higher without altering the structure of the rules, the percentage of incorrect classifications dropped to 7%.

With these boundary changes made, data from three new speakers, two male and one female, were used to test the rules. The score for correct and unique recognition dropped to 78.5%. The pooled formant energy difference data for the four speakers are found in Figures 3.2(a)-(e). Again data where parameter measurement accuracy was doubtful were excluded. Several significant points emerged:

(1) Rules concerned with the parameters Duration(dip(S)) and Duration(dip(R)) were not useful in distinguishing nasals from liquids for any speakers other than the original speaker. Thus it was decided to scrap the rules involving these parameters, which left the formidable task of finding new distinguishing rules for the cases /a/-consonant-vowel and /i/-consonant-vowel.

(2) Points which had seemed to be oddly isolated for the original speaker's data were found to be clustered with other sounds of the same type when the data for the four speakers were pooled. This can be seen particularly on the /a/-consonant-vowel and /i/-consonant-vowel graphs. (Figures 3.2(a) and (c)).

(3) The two liquid sounds /l/ and /r/ seemed to have their formant energy difference parameters in different regions, i.e. instead of finding clusters of liquids one tends to find clusters of /l/'s and clusters of /r/'s.

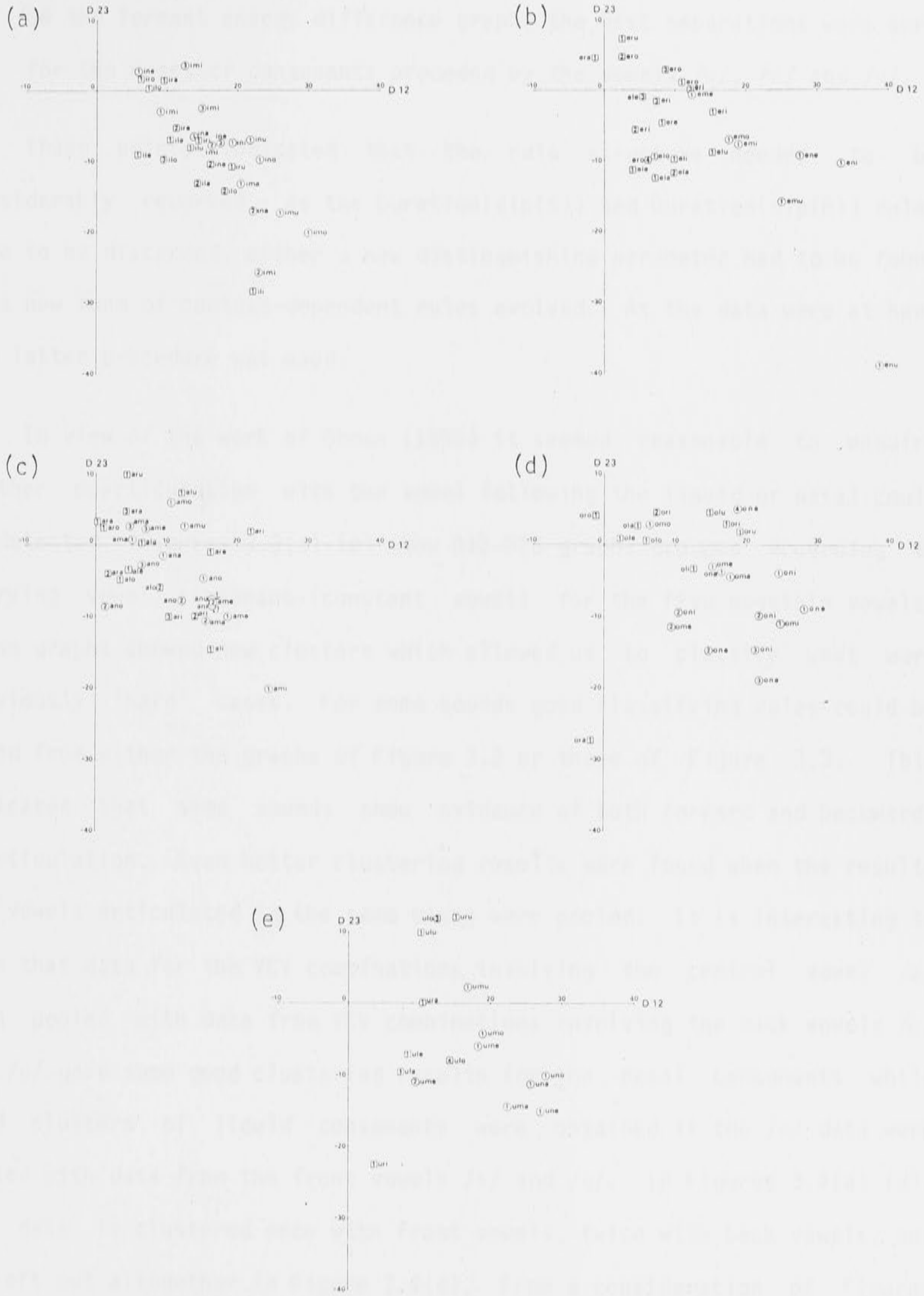


FIGURE 3.2(a)-(e): D23 versus D12 for the cases X-consonant-varying vowel, where X stands for the vowel /i/, /e/, /a/, /o/, or /u/ in Figures 3.?(a), (b), (c), (d), and (e) respectively. Data for four speakers.

(4) On the formant energy difference graphs the best separations were seen for the cases of consonants preceded by the vowels /e/, /o/ and /u/.

These points indicated that the rule structure needed to be considerably reworked. As the Duration(dip(S)) and Duration(dip(R)) rules were to be discarded, either a new distinguishing parameter had to be found or a new form of context-dependent rules evolved. As the data were at hand the latter procedure was used.

In view of the work of Ohman (1966) it seemed reasonable to enquire whether coarticulation with the vowel following the liquid or nasal could be detected. Figures 3.3(a)-(e) show D12-D23 graphs grouped according to (varying vowel)-consonant-(constant vowel) for the five possible vowels. These graphs showed new clusters which allowed us to classify what were previously 'hard' cases. For some sounds good classifying rules could be found from either the graphs of Figure 3.2 or those of Figure 3.3. This indicates that some sounds show evidence of both forward and backwards coarticulation. Even better clustering results were found when the results for vowels articulated in the same place were pooled. It is interesting to note that data for the VCV combinations involving the central vowel /a/ when pooled with data from VCV combinations involving the back vowels /o/ and /u/ gave some good clustering results for the nasal consonants while good clusters of liquid consonants were obtained if the /a/ data were pooled with data from the front vowels /i/ and /e/. In Figures 3.4(a)-(d), /a/ data is clustered once with front vowels, twice with back vowels, and is left out altogether in Figure 3.4(d). From a consideration of Figures 3.2, 3.3 and 3.4 the following rules were developed using new fuzzy restrictions or D12 and D23.

FIGURE 3.3(a)-(e): D23 versus D12 for the cases varying vowel-consonant-X, where X stands for the vowel /i/, /e/, /a/, /o/, or /u/ in Figures 3.3(a), (b), (c), (d), and (e) respectively. Data for four speakers.

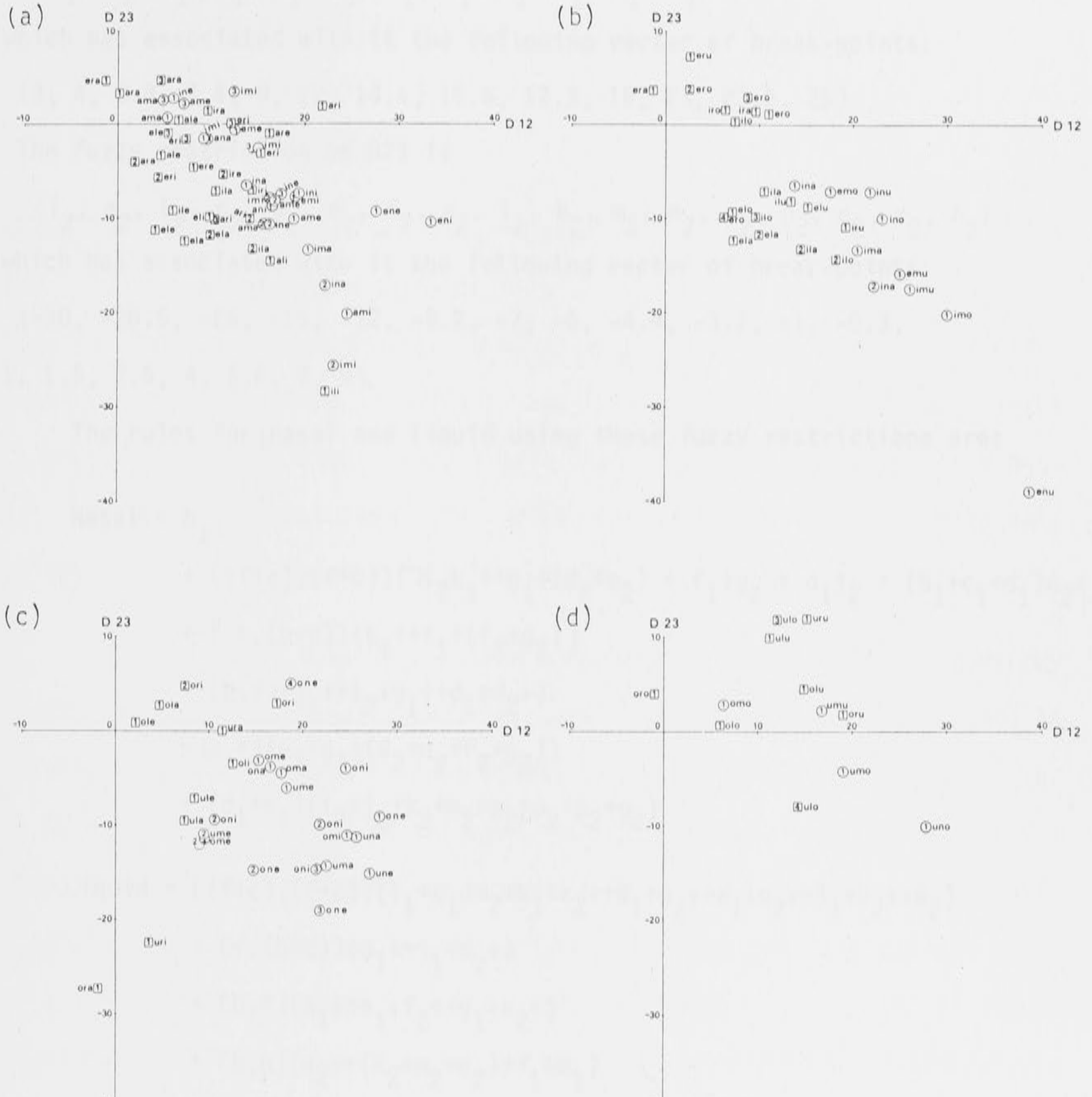


FIGURE 3.4(a): D23 versus D12 for the case (front or central vowel)-consonant-(front or central vowel)
3.4(b): D23 versus D12 for the case (front vowel)-consonant-(central or back vowel)
3.4(c): D23 versus D12 for the case (back vowel)-consonant-(front or central vowel)
3.4(d): D23 versus D12 for the case (back vowel)-consonant-(back vowel)

Note: Data for four speakers in all cases.

The fuzzy restriction on D12 is:

$\{l_1, a_1, b_1, c_1, d_1, e_1, f_1, g_1, i_1, j_1, k_1, h_1\}$

which has associated with it the following vector of break-points:

{3, 4, 5.3, 7.4, 9, 12, 14.4, 15.6, 17.3, 19, 23, 23.5, 25}

The fuzzy restriction on D23 is

$\{l_2, a_2, b_2, c_2, d_2, e_2, f_2, g_2, i_2, k_2, m_2, n_2, o_2, p_2, q_2, r_2, h_2\}$

which has associated with it the following vector of break-points:

{-30, -28.5, -26, -15, -12, -8.2, -7, -6, -4.4, -3.2, -1, -0.3,
1, 1.5, 2.6, 4, 5.6, 7, 9}

The rules for nasal and liquid using these fuzzy restrictions are:

$$\begin{aligned}\text{Nasal} = {} & h_1 \\ & + [(f+c),(f+c)](\neg l_2 k_1{\uparrow} + g_1{\uparrow}(d_2+e_2) + f_1{\uparrow}g_2 + d_1 j_2 + (b_1+c_1+d_1)n_2{\uparrow} \\ & + [f,(b+c)](k_1{\uparrow} + f_1{\uparrow}(f_2+g_2)) \\ & + [b,f](j_1{\uparrow} + i_2{\downarrow}g_1{\uparrow} + d_1{\uparrow}d_2{\downarrow}) \\ & + [a,*](e_2+g_1{\uparrow}(d_2+e_2+f_2+g_2)) \\ & + (d_1+e_1)(i_2+j_2+k_2+m_2+n_2+o_2+p_2+q_2)\end{aligned}$$

$$\begin{aligned}\text{Liquid} = {} & [(f+c),(f+c)](l_1+c_1{\uparrow}m_2+b_1{\uparrow}k_2{\downarrow}+d_1{\uparrow}g_2{\downarrow}+e_1{\uparrow}e_2{\downarrow}+j_1{\uparrow}b_2{\downarrow}+e_2) \\ & + [f,(b+c)](d_1{\downarrow}+i_1{\downarrow}d_2{\downarrow}) \\ & + [b,f](b_1{\downarrow}+e_1{\downarrow}f_2{\uparrow}+g_1{\downarrow}k_2{\uparrow}) \\ & + [b,b](q_2{\uparrow}+(k_2+m_2+n_2)+f_1{\downarrow}m_2) \\ & + [c,b](h_2+l_1 g_2{\uparrow}+c_1{\downarrow}g_2)\end{aligned}$$

It is also possible to consider rules for /ɹ/ and /ɾ/:

$$\begin{aligned}/ɹ/ = {} & [(f+c),(f+c)](a_1(j_2+k_2+m_2) + c_1{\downarrow}e_2{\downarrow} + \neg k_1 b_2{\downarrow}) \\ & + [f,(b+c)]((c_1+d_1)k_2{\downarrow}+(c_1+d_1+f_1)d_2{\downarrow}) \\ & + [b,f](l_1 d_2{\uparrow}+b_1 d_2+d_1(h_1+g_2+i_2)) \\ & + [b,b](e_1 q_2{\uparrow}+a_1 n_2+e_1 d_2) \\ & + [c,b](e_1{\uparrow}h_2)+(a_1+b_1+c_1)g_2)\end{aligned}$$

$$/r/ = [(f+c),(f+c)](l_1 g_2{\uparrow}+g_1{\uparrow} i_2{\uparrow}+d_1 m_2+d_1{\downarrow} g_2+f_1(d_2+b_2+f_2) + e_1(d_2+e_2)$$
$$+ [f,(b+c)](o_2{\uparrow}+a_1{\uparrow}+i_1 d_2{\downarrow})$$
$$+ [b,f](a_1{\downarrow} c_2{\downarrow}+(c_1+d_1+e_1+f_1+g_1)k_2{\uparrow})$$
$$+ [b,b](l_1+i_1 n_2+g_1{\uparrow} h_2)$$
$$+ [c,b](c_1\ h_2+\overline{\ }a_1{\uparrow} c_2{\uparrow})$$

where '↑' stands for 'and anything greater'

and '↓' stands for 'and anything less'.

Thus '$f_1{\uparrow}$' is shorthand for '$(f_1+g_1+i_1+j_1+k_1+h_1)$'. f, c, b, as used for vowel contexts, stand for front vowel, central vowel and back vowel respectively and all other symbols have their previously defined meanings. It can be seen that separate rules for /l/ and for /r/ involve only slightly more work than the rule for liquid.

For cases where despite grouping according to various different nearby vowels, a cluster of nasal sounds is close to a cluster of liquid sounds in D12-D23 space, the sounds in this cluster although they will be recognised as nasals, will be placed in the nasal category with only slightly over 0.5 membership (and will probably be placed in the liquid category with only slightly less than 0.5 membership). This indicates that the D12-D23 space separation of these sounds is not particularly strong. In spite of these close-to-borderline cases the recognition rules have proved satisfactory in distinguishing liquid and nasal sounds with 95% accuracy, and in distinguishing /l/ from /r/ with 88% accuracy for many speakers of Italian*.

* It has recently been found from very extensive tests that a simplified version of these rules can be used to achieve substantially the same accuracy (De Mori, 1980).

3.3 DISCUSSION

A recognition algorithm for distinguishing between the liquid consonants (either as a single class or as two separate classes /l/ and /r/) and the nasal consonants has been presented. It was found that the two parameters, the difference between the energy associated with the first and second formants (D12) and the difference in the energy associated with the second and third formants (D23)* gave good separations between the classes if the results were grouped according to the vowels adjacent to the sounds under consideration. That class separations should be improved if results are grouped according to the identity of the neighbouring vowels fits in with Fant's notion of relational invariance of the manifestations of a distinctive feature (Fant, 1967). This would imply that a change in a parameter in going to a particular sound will not only depend on the type of movements made by the articulatory muscles involved but it will also depend on the initial position of the articulators.

As a general result for the nasal coarticulations, D12 was large and positive. Also in the case of nasals the energy associated with the third formant is greater than that associated with the second formant. A minor exception to these rules is the class of (front vowel)-(nasal consonant)-(front vowel) combinations. For almost fifty percent of the sounds in this class, and particularly for the nasal /m/, the energy associated with the first formant is only slightly greater than the energy associated with the second formant and the second and third formants have each about the same amount of energy. A related effect was noted by Fujimura (1962) who found that the zero associated with /m/ rises in

* This is a somewhat loose use of the term 'formant' as, correctly speaking, only vowels have formants. Also should be noted that formants for nasal consonants may not always be well defined (because of damping and cancellation by zeros), and consequently measures that require identification of formants may be subject to a lot of variability.

frequency when /m/ occurs before a front vowel, while the position of the second formant stays relatively constant. Thus less pole-zero cross-interaction would be expected in the case of /m/ before a front vowel than in the case of /m/ before a central or back vowel.

For /r/ and /l/ D12 is generally not great; the energy associated with the first formant, except for 6% of the vowel-/r/-vowel cases, is greater than the energy associated with the second formant. For the case of a liquid occurring between a front or central vowel and a central or back vowel the energy associated with the third formant is generally less than the energy associated with the second formant for /r/ but greater for /l/. Indeed in 71% of all the considered coarticulations with /l/ the energy associated with the third formant is greater than the energy associated with the second formant.

The approach of not using more than two parameters might be questioned. Our justification for this is that in designing a component of an automatic Speech Understanding System, the saving of processing time is important. As vowel recognition is the first step after pre-categorical classification, information about vowels is already available and the D12 and D23 parameters are easily and quickly obtained. Examination of the data indicated that the fuzzy rules show by the answers they yield where the errors in the rules lie. Should more extensive testing show need for some rule modification, it is easily made because the liquid/nasal distinguishing program has been written to be completely general, so that only the input data structure would have to be changed.

Perhaps the most significant aspect of the rules is their speaker independence. This is particularly interesting in the light of a discussion by Mermelstein (1977) on various methods that have been evolved

for nasal consonant identification. He states that 'interspeaker variation appears the most significant limitation to improvement of the classification results.' Also he notes that nasals are often confused with liquids before high vowels. In the algorithm described here the liquid/nasal distinction before the two high vowels /i/ and /u/ is very robust and achieves high fuzzy membership scores in the appropriate categories. The good classification achieved relatively simply here suggests that good acoustic-phonetic classifcation can be achieved by developing suitable contextual rules. This would lead to reduced complexity of higher level processing in Speech Understanding and Speech Recognition Systems.

One thing has not been tested: whether the system will outclassify bad cases, i.e. when a non-liquid, non-nasal sound is presented to it, will it classify the sound with membership less than 0.5 in both the liquid and the nasal categories? This is desirable in order that errors at the pre-categorical classification stage not affect results at higher levels. It is intended that testing for this feature and appropriate rule modification will be the next stage in this work.

Chapter 4

COMPUTER RECOGNITION OF PLOSIVE SOUNDS USING CONTEXTUAL INFORMATION

4.1 INTRODUCTION

The objective of the work in this chapter is to find an effective and comprehensive method of automatically recognizing the plosive consonants /p, b, t, d, k, g/. As plosive consonants have been the subject of very extensive research by psychologists and phoneticians, the mechanisms of production and perception of these consonants are well documented. The system proposed in this paper is based on this knowledge to ensure that the automatic recognition rules imitate, as closely as possible, the processing of the human brain at the acoustic-phonetic level. In Section 4.2 past research on the plosive consonants is reviewed and thus features which are essential in a plosive recognition scheme are highlighted. Section 4.3 is a discussion of existing plosive recognition schemes. In Section 4.4 a philosophy of plosive recognition rule development for Italian plosives is presented. In Section 4.5 the rules for a precategorical classification necessary for isolating plosive segments are described; in Section 4.6 some experimental results are described*.

* As mentioned in the Statement (p.ii), this chapter was the result of joint work. The author's contribution was the literature survey and the development of the initial version of the recognition rules based on data for one speaker.

4.2 REVIEW OF RESEARCH CONCERNING THE PLOSIVE CONSONANTS

This section gives an historical overview of some of the diverse forms of plosive research, to show whence the motivation for choosing the parameters used in this approach were derived.

Psychological studies on speech perception are of two types: those that use synthetically generated speech and those that use real speech. The synthetic speech experiments have highlighted the fundamental characteristics of plosives while the real speech experiments have tended to clarify the interaction of these characteristics and the variability of their realizations in real speech.

In synthetic speech experiments in the early 1950's Cooper, Delattre, Liberman, Borst, and Gerstman (1952) found that certain stop burst spectra were characteristic of the three stop types. In later experiments (Liberman, Delattre and Cooper, 1952) they found that the formant transitions from the plosive consonant to the following vowel produced a successful synthetic voiced plosive. The first formant transition appeared to contribute to the voicing of the stop while the second formant transition provided a basis for distinguishing between the stop types. Further experiments led to development of the now famous 'locus theory' which postulates that the second formant transitions should point to a frequency locus no matter what the following vowel is. Delattre, Liberman, and Cooper (1955) found that this was particularly characteristic of /d/, not quite so reliable for /b/, while for /g/ there were two loci, a high one if the following vowel was a front vowel and a low one if the following vowel was a back vowel. Attempts to locate third formant loci were also made but these were not conclusive although it was obvious that the third formant transitions could be an important secondary plosive cue (Harris et

al., 1958). Hoffman (1958) further refined much of the previous plosive research on stop consonants to see how the two main types of cues for plosive consonants, burst and formant transitions, interacted. He concluded that all the cues are perceptually independent of the other cues present and that for some stops, notably /b/ and /p/, the burst provided a weak cue and the transitions a strong cue while for /d/ the burst was a strong cue and the transitions were a weak cue. Later research has demonstrated the significance of these conclusions.

In real speech experiments, Miller and Nicely (1954) had listeners identify nonsense syllables spoken by five female speakers over a voice communication system which was characterised by frequency distortion and random masking noise. In this way they decided that the most robust cue to the place of articulation of the voiced plosives was the initial portion of the vowel that followed; while for the voiceless plosives the burst was more characteristic as the transitional portion was somewhat masked by the period of aspiration.

In further real speech experiments, Halle et al. (1957) considered plosives occurring not only in conjunction with vowels but also in consonant clusters and at the beginnings and ends of words. They concluded that a complex array of cues was needed to characterize the plosives and that the locus theory was somewhat inadequate for this task. In this they were supported by Ohman (1966) who studied spectra of VCV coarticulations for Swedish, using all possible combinations in which the consonant was a plosive. He deduced that each VCV coarticulation was a 'basic dipthongal gesture with an independent stop consonant gesture superimposed on its transitional portion'. By the end of the 1960's there was no significant agreement on what provided the best perceptual cue for classification

within the class of plosives. Each researcher had excellent experimental material to support his view as to whether burst or transitions was more significant. This dichotomy is still unresolved. The burst-transition debate continued with the publication of papers by Sharf and Hemeyer (1972), Wintz, Scheib, and Reeds (1972), and La Riviere, Wintz, and Herriman (1975). The bulk of the controversy is summarized and elaborated by Cole and Scott (1974a,b). While acknowledging that a diversity of cues is useful for identification of stops they argued that all speech sounds can be considered to consist of one important basic gesture. In the case of the plosives this is the burst. The transitions are considered to be just the movements between the characteristic invariant positions of two phonemes*.

This view is challenged by Dorman, Studdert-Kennedy and Raphael (1977) CVC nonsense syllables spoken by two speakers were presented to listeners for identification after certain portions of the syllables had been excised or jumbled. It was found that results were consistent for each speaker but that results differed markedly between the two speakers. One speaker's plosives, particularly for the back vowel-plosive coarticulations, could be identified from the burst portion of his CVC coarticulations but for the other speaker the burst cues were always weak while the transition cues were strong. This gives a very salutary warning about the dangers of averaging results for different speakers.

* This view has been supported to some extent by recent experiments by Stevens and Blumstein (1978; Blumstein and Stevens, 1979, 1980). These papers are discussed extensively in Chapter 6 where it is shown that the Stevens and Blumstein approach to burst recognition can be used to simplify the automatic recognition of plosive consonants.

In all the works considered so far, we have seen that results about plosive consonants are invariably presented with reference to the sounds produced in conjunction with the plosive. Ohman's investigation (1966) of these coarticulatory effects has already been mentioned. Using conventional cineflurographic techniques, Gay (1972,1977) has studied coarticulation in VCV sequences where the vowels used were /p,t,k/ and the consonants /i,a,u/. He concluded that both anticipatory and carryover coarticulation effects were present but that the coarticulation mechanism operated only between neighbouring sounds and was therefore much simpler than the 'diphthongal gesture' proposed by Ohman. Gay also found that carryover coarticulation depended on the phonetic identity of the phone on which it might act but that no corresponding result existed for anticipatory coarticulation. Another examination of coarticulation effects was carried out by Butcher and Weiher using electropalatography (1976). They too noticed anticipatory and carryover coarticulatory effects; the strongest coarticulatory effects occurring with the vowel /i/. Also they found some evidence for coarticulatory effects extending from one vowel to the other across the plosive consonant in VCV sequences. Perhaps their most important discovery is that there are wide interindividual differences in the extent and type of coarticulation.

Leaving the articulatory mechanisms and turning to the mechanisms of perception, we are faced with the problem of how the signal received by the ear is processed and ultimately recognized by the brain. This is of particular importance in the case of the plosives as they are characterized by several features. Research with adaptation followed by testing paradigms indicates that there exist feature detectors at some level (Eimas and Miller, 1972), i.e. detectors that in some as yet unknown way respond to events such as 'burst' and 'transition'. Furthermore, Miller and Eimas

(1976) have shown that feature detectors can be tuned. Such tuning would account for the manner in which slight differences in articulation as say in the production of /t/ in Italian and /t/ in English can be perceived after sufficient training. The mode of integration of the information from the feature detectors is not known with any certainty. A possible explanation is given in a model presented by Oden (1978). He proposes that each feature is fuzzily identified as belonging to one of several classes. After this, there is prototype matching in which fuzzy identification rules for each phoneme are applied to the results output by the feature detectors. Finally the phoneme is classified as the phoneme whose composite feature detection pattern is most closely matched by the composite feature detection pattern of the test phoneme.

In this section it has been shown that the plosives are characterized by several features: transitions, bursts and timing. It seems that for each plosive there exist several sufficient but not necessary combinations of appropriate features. It has also been shown that the articulation of plosives is influenced by the sounds produced in conjunction with them. The perception mechanism consists of feature detectors, the information from which is integrated to produce identification under the control of rules that may be nondeterministic and have to account for the different combinations in which features have been detected.

4.3 EXISTING AUTOMATIC PLOSIVE RECOGNITION SCHEMES

Here some of the existing automatic plosive recognition systems are reviewed. Such a review is enlightening as it reveals which plosive characteristics are particularly amenable to automatic location and quantitative measurement. Five of the nine schemes considered here are systems in which the class of plosives was systematically investigated with

a view to finding parameters which would accurately yield plosive identification irrespective of machine time taken and the constraints of working within a large scale recognition system. These are the systems of Halle et al., (1957); Datta, Ganguli, Ray, and Mukherjee (1978); Alinat (1978); Searle et al. (1979); and Fujisaki, Tanaka, and Higuchi (1979). The other four systems are ARPA systems in which the constraints of multi-level processing and limited overall development time dictated compromises which yield fast and reasonably accurate, but by no means perfect, recognition at the acoustic-phonetic level.

The work of Halle et al. (1957) has already been mentioned. The heirarchy of their recognition system has been adopted in almost all later recognition schemes. When considering ways to locate a plosive in continuous speech they point out that the production of a plosive requires several centiseconds during which there is no energy except for a possible voicing component. Transitions and bursts may or may not be perceived depending on the context. As regards the voiced/unvoiced decision Halle et al. point out that the essential distinction between /b/,/d/,/g/ and /p/,/t/,/k/ in English is that in 'the production of the the latter more pressure is built up behind the closure than in the production of the former'. Thus they prefer that the distinction be called a tense/lax decision. Having found that suitably trained subjects could learn to identify aurally individual plosives when the burst portion alone was presented to them, Halle et al. concluded that the burst was a sufficient cue and proceeded to develop a burst recognition scheme. This scheme consisted of first measuring the intensity in the 700-10,000 Hz and 2,700-10,000 Hz ranges. Those sounds with significant energies in the higher frequency range were /t/,/d/ and /k/,/g/ before front vowels. These sounds were labelled 'acute'. The 'grave' class consisted of /p/,/b/ and

/k/,/g/ before back vowels. This grave class was then subdivided into two subclasses by measuring the frequency position of the highest spectral peak and the acute class was subdivided on the basis of further energy band measurements. Up to 85% correct recognition was obtained by this method.

Halle et al. also attempted to develop rules for plosive recognition using the transitions. An initial problem here arises when one wishes to accurately specify the transition in position and time. Nevertheless gradient rules were found which while they were consistent with the Haskins Laboratory locus theory were considerably more complex than had been suggested by the Haskins research.

The work by Halle et al. remained the standard work on plosive recognition rules until the publication of the results of the ARPA projects. The first of the ARPA systems to be considered here is the work carried out by Weinstein et al. (1975) at the Lincoln Laboratories, MIT. This work is notable among the ARPA projects because of its effort to provide very accurate acoustic-phonetic data. To train the system, male and female voices were used and the recordings were done in a terminal room atmosphere. One hundred and eleven sentences were used - these sentences in general concerned command of a speech data base. No attempt was made to achieve phonetic balance. Careful initial segmentation and classification were carried out, and during this process plosives were located as a conjunction of a silence with or without voice bar, a plosive burst (if present) and aspiration (if present). The duration of the silence following the burst is critical to plosive detection - a silence duration exceeding 70 msec means that the sound is tagged as a fricative. A voiced/unvoiced decision was made according to the output of the pitch detector. However only a small number of the /b/,/d/,/g/ sounds were

tagged as voiced although no /p/,/t/,/k/ sounds were ever so tagged. Recognition according to place of articulation was done by 'finding the frequency location and the relative strength of the major concentrations of energy in the burst spectrum'. When an extremely low main energy concentration frequency is found it is concluded that the burst location is incorrect. An algorithm to detect post-vocalic /k/ and /g/ from formant transitions into the pre-burst silence was also developed. It was felt that recognition rates would have been improved if the rules had been made speaker dependent.

Another ARPA project in which quite a lot of attention was paid to plosive recognition is described by Woods (1976). Acoustic-phonetic rules were initially developed as the result of noting the performance of speech researchers in parameter reading sessions. Energy band parameters were used for initial classification into broad phonetic classes. Thus voiced plosives were tagged by a characteristic dip in the overall energy. The voiced/unvoiced decision was made by a voice onset time (VOT) measurement. For plosives occurring before vowels a two-pole frequency approximation to the peak for the 20 msec analysis window centred on the burst was used for classification according to place of articulation. An auxiliary classificatory measurement was the change in the third formant just before the silence. To improve classification it was later decided that plosive allophone algorithms should be developed particularly for cases of plosives followed by /r/. Through each stage of the recognition probabilistic scoring was employed.

Two other ARPA projects were top-down in strategy. These are the SRI project the acoustic-phonetic component of which is described by Becker and Poza (1975) and the Carnegie-Mellon Harpy system whose segmentation and

labelling procedure is described by Goldberg (1975). In neither of these systems is particular emphasis given to a plosive recognition component per se. The SRI system takes a dual approach to recognition at the acoustic-phonetic level. What might be called 'conventional' analysis is done insofar as the output to four pass-band filters is used to output a sequence of labels indicating to which of ten classes each speech segment belongs. After this, word verification routines involving selected formant tracking were used to recognize the utterance. Confidence levels can be determined during the application of these routines. The second part of the approach is to verify hypotheses using analysis-by-synthesis techniques.

The Carnegie-Mellon Harpy system was based on template-matching. Labelling and segmentation was done by making a particular set of measurements and then seeing how closely they agree with the measurements of a standard set. Many distance metrics are employed and probabilistic scoring was used.

Recently, Datta et al. (1978) derived a system for plosive recognition from examination of a data set of 600 plosive-vowel combinations spoken by three male speakers of Telagu. The plosive sounds of Telagu differ from those of English in that they are of four different articulatory place types - labial, dental, alveolar, and velar - and that they have no associated aspiration. The three features used for classification were the first and second formant transitions and the duration of these transitions. The transition measurements were made by extrapolating the formants back to the point of release of the burst by eye and then measuring the difference between these points and the formant positions during the steady state portion of the vowel. A

maximum-likelihood classification was used, based on the three measurements described. It was found that the best classification occurred when the plosives had been separated a priori into voiced and unvoiced sets and the target vowels were not pooled. Very good recognition scores were reported.

In contrast to the system of Datta et al. (1978) in which transitions and timing were used for classification Alinat (1978) developed a system which relied heavily on measurements of the burst to achieve good classification. An artificial cochlea was used to analyse all CV syllables composed of the six French plosives and nine selected French vowels. The speaker group comprised twenty-four male speakers and ten female speakers. Plosives (followed by vowels) were located by the distinctive shape of the energy curve. To classify the sounds according to place the parameters used were duration of the burst, ratio of high energy to low energy, and four geometric measurements on the evolution of the energy curve. Rules have been developed using these parameters. The /p/,/b/ versus /t/,/d/ distinction was made on what was essentially a centre of gravity measurement while /k/,/g/ were characterized by the burst spectrum with a foreknowledge of the second formant of the following vowel. Confusion matrices given show that the classification was better than 82% successful in all cases.

Searle et al. (1979) have also developed a system for discrimination of stop consonants based on studies of auditory physiology. Their system consists of a one-third octave filter bank which models the auditory tuning curves, and a bank of high-speed, wide dynamic range envelope detectors. The outputs from the filter detector channels are used to obtain information about various plosive features such as voice onset time, location and shape of the burst, location and shape of the spectral peaks

during the transitions, and the slopes of the transitions. A discriminant analysis programme is used to classify the data. Performance of a test set of 148 stop consonant tokens was 77%.

Fujisaki et al. (1979) have developed a system for classification of the voiced stop consonants in which formant frequencies of the transitions from the stop consonant to the following vowel are very carefully tracked using bandwidth and continuity constraints as aids to finding the correct trajectory of each formant. Second and third formant loci are then estimated and it was found that the values of these two parameters were sufficient to give good separations between the three types of voiced stop consonant.

4.4 AN ALGORITHM FOR THE RECOGNITION OF PLOSIVE CONSONANTS

4.4.1 Overview

The algorithm for the recognition of the plosive consonants is a fuzzy one because it assigns a degree of compatibility between an acoustic pattern p and an hypothesis H_p. H_p can be seen as a label assumed by a linguistic variable taking values on the set P of the plosive sounds:

$$P = P_T UP_L,$$

where:

$P_T = \{p, t, k\}$:set of the plosive tense sounds;

$P_L = \{b, d, g\}$:set of the plosive lax sounds.

This fuzzy algorithm is executed whenever a consonant is hypothesised in an interval (t_i, t_j) of the acoustic pattern and the possibility that the consonant may be nonsonorant and interrupted is not negligible.

This possibility is evaluated by fixing thresholds on the membership of the features 'sonorant' and 'interrupted' in order to ensure that the probability of having at least one membership in these classes below the threshold is less than 0.1 for sounds that are known to be nonsonorant interrupted consonants.

The fuzzy algorithm for the recognition of plosive consonants is further split into two parts, one for the lax plosives devoted to the generation of hypotheses in P_L and one for the tense plosives for the generation of hypotheses in P_T.

Again, the execution of one of the two parts of such an algorithm depends on which of the features 'lax' or 'tense' is obtained after another fuzzy algorithm is applied to those sounds for which the presence of the feature 'nonsonorant-interrupted' has been hypothesised previously.

The generation of hypotheses in P_L will be considered here for describing this methodological approach. Hypothesis generation is performed by a fuzzy algorithm based on a set of fuzzy composite questions:

$Q_{HPL} \triangleq$ ' is HPL in $p\{t_i,t_j\}$?'

where HPL is a variable taking values in P_L, and $p\{t_i,t_j\}$ is an acoustic pattern obtained by taking the segment (t_i,t_j) of the whole pattern p of the speech signal to be interpreted.

The aim of the fuzzy algorithm is to compute the possibility:

Poss (HPL is in $p\{t_i,t_j\}$) V HPL $\in (P_L)$

HPL is a complex concept defined by the composite question Q_{HPL}. As was discussed in Chapter 2, a composite question defining a complex imprecise concept is a triple:

$$Q \triangleq (U,R,A)$$

where U, the universe of discourse, is in this case the set of all possible patterns; B, a structured linguistic variable is in this case a value, such as /b/, that may be assumed by HPl, and A is a set of possible answers, in this case a set of evidence statements which may be deduced from Poss HPL is in $p\{t_i,t_j\}$.

4.4.2 Definition of the Terminal Alphabet for the Recognition of Plosive Consonants

For the recognition of plosive consonants two types of acoustic features have been considered, namely formant transitions (FOR) and burst spectra (SP).

For each plosive sound and for each of the two features just introduced a nonterminal symbol has been introduced in V_N the non-terminal alphabet of acoustic features.

These symbols are defined in Table 4.1.

Let X denote a generic plosive sound, XFOR and XSP are expressed as functions of some components which are represented by other variables in V_N. The rules and the symbol used depend on the type of syllabic nucleus in which the stop sound appears. Syllabic nuclei of the type VCV for lax plosives and CV for tense plosives are considered in this chapter. It is believed that most of the rules presented here can be generalised and applied to plosives in other types of syllabic nuclei.

BFOR	formant transitions for /b/
DFOR	formant transitions for /d/
GFOR	formant transitions for /g/
PFOR	formant transitions for /p/
TFOR	formant transitions for /t/
KFOR	formant transitions for /k/
BSP	burst spectrum for /b/
DSP	burst spectrum for /d/
GSP	burst spectrum for /g/
PSP	burst spectrum for /p/
TSP	burst spectrum for /t/
KSP	burst spectrum for /k/

TABLE 4.1: Symbols associated with transitions and burst features of plosive consonants

The syntactic category XFOR is expressed in terms of the following features which are represented by symbols in V_N:

XFL : formant pseudo-loci

XFS : formant slopes

XBZ : buzz-bar characteristics at the onset of the first formant toward the vowel following the plosive.

XBZ appears only in the rules of XFOR with $X \in P_L$. When $X \in P_L$, XFL and XFS are expressed in terms of two other nonterminal symbols, namely XFLP, XFLF XFSP and XFSF; where the last letter makes reference to the vowel preceding the plosive (P) or following the vowel (F).

The general rule for rewriting the phonemic hypothesis $X \in V_N$ is:

$$X \overset{\alpha_X}{\rightarrow} XFOR$$

$$X \overset{\beta_X}{\rightarrow} XSP$$

$$X \overset{\gamma_X}{\rightarrow} XFOR.XSP$$

Where the α_X, β_X, $\gamma_X \in [0,1]$ represent the 'grammaticalities' of the corresponding syntactic rules. As XFOR and XSP will be hypothesised with degrees of compatibility with the pattern $p \in U$ to be analysed, the following semantic rule belonging to S_N is used for obtaining the degree of compatibility (the possibility of similarity) between p and the hypothesis X:

$$\text{Poss}(X \text{ is in } p) = \alpha_X \wedge \text{Poss}(XFOR \text{ is in } p\ V)$$
$$\vee \beta_X \wedge \text{Poss}(XSP \text{ is in } p) \vee \gamma_X \wedge \text{Poss}(XFOR \text{ is in } p)$$
$$\wedge \text{Poss}(XSP \text{ is in } p)$$

All the possibilities are membership values belonging to the [0,1] interval. For the sake of simplicity, the syntactic rules defining X are expressed in algebraic form as follows:

$$X = \alpha_X XFOR + \beta_X XSP + \gamma_X XFOR.XSP$$

In order to define the symbols of the terminal alphabet a training set consisting of a collection of 100 VCV syllables pronounced by one male speaker and containing all possible coarticulation instances was considered. The syllables were extracted from continuous speech and segmented automatically by an algorithm described by De Mori et al. (1976). After segmentation, formants were tracked using an algorithm described by Laface (1980) and the formant psuedo-loci were extracted.

The pseudo-loci are the values of the formant frequencies taken when the formant amplitude has an abrupt decay (before the plosive) or a large increase at the transition toward the vowel following the plosive. These values are indicated as FiY where i = (1,2,3) indicates the i-th formant, Y is P or F depending on whether the psuedo loci are taken at the beginning or at the end of the plosive respectively. From the data collected, some broad 'fields of existence' of the plosive sounds in different contexts were identified and represented on the diagrams of Figures 4.1(a) and 4.1(b).

It is important to note that these fields have been drawn on the basis of some experimental data and on EXPECTATION from knowledge and experience of the phonetic models for the production of plosive sounds. (The experiment described in Chapter 6 verifies the substantial correctness of this expectation.)

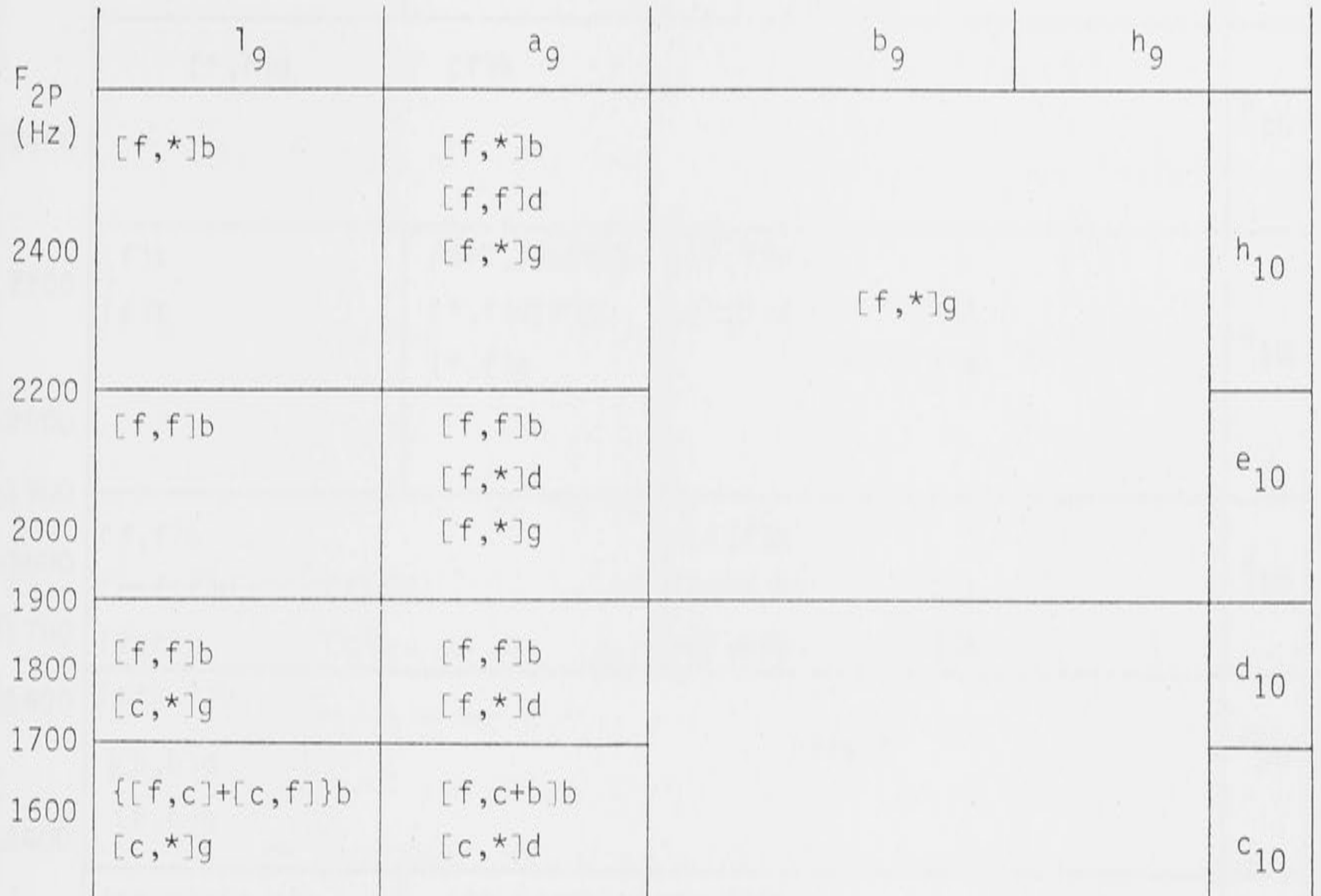


FIGURE 4.1(a): Fields of existence in F2-F3 space of the endpoints of transitions from a vowel to the plosive consonant following it. The contents of the square brackets indicate the vocalic context. Fuzzy interval labels are marked along the top and right side of the figure.

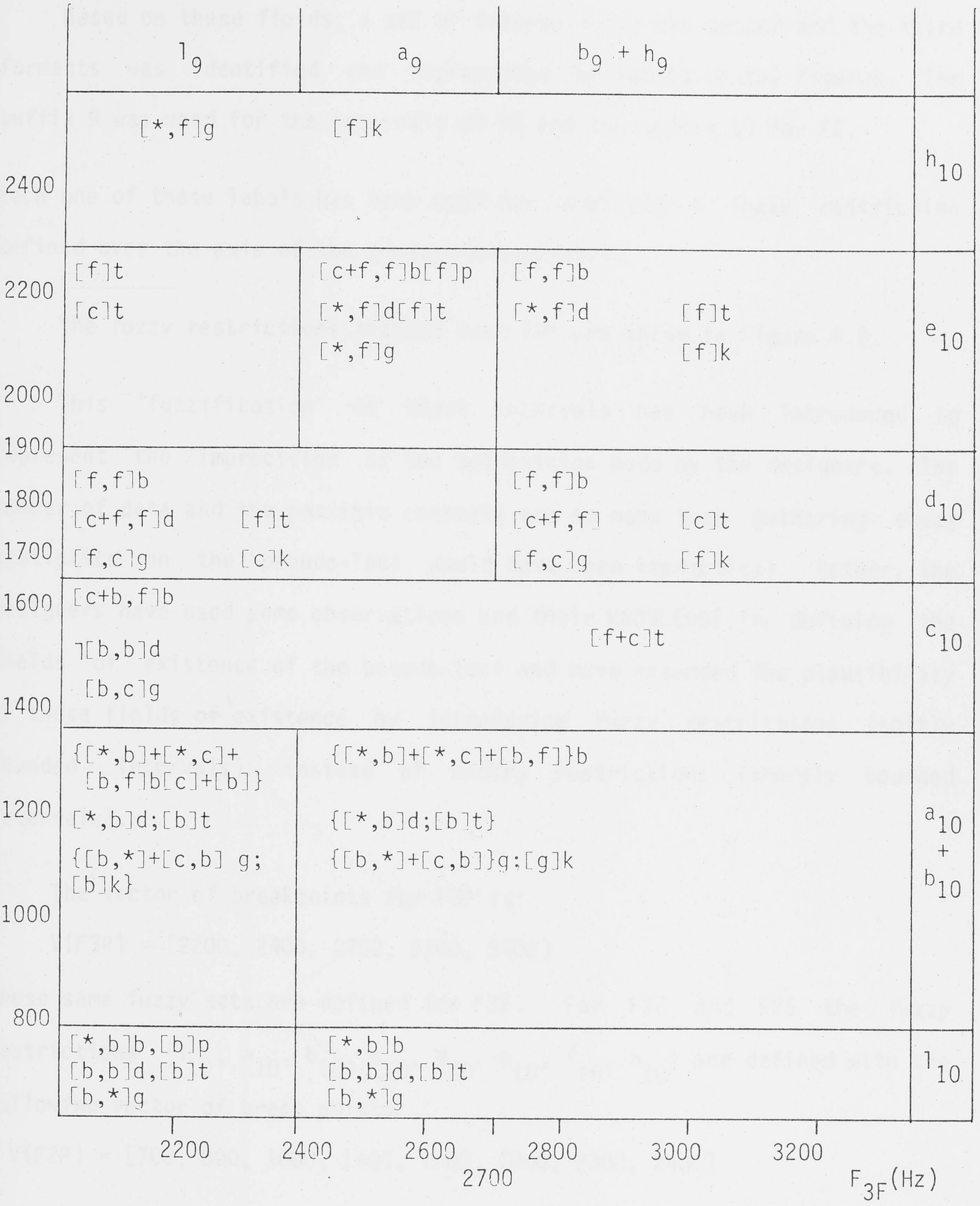


FIGURE 4.1(b): Fields of existence in F2-F3 space of the beginning-points of transitions from a plosive consonant to the vowel following it. The contents of the square brackets indicate the vocalic context. Fuzzy interval labels are marked along the top and right side of the figure.

Based on these fields, a set of intervals for the second and the third formants was identified and represented by labels on the figures. The suffix 9 was used for the intervals of F3 and the suffix 10 for F2.

Each one of these labels has been used for labelling a fuzzy restriction defined over the axis of the formant frequencies.

The fuzzy restrictions defined over F3P are shown in Figure 4.2.

This 'fuzzification' of these intervals has been introduced to represent the imprecision of the subdivision made by the designers. The number of data and the possible contexts are so many that gathering exact statistics on the pseudo-loci would have been impractical. Rather, the designers have used some observations and their KNOWLEDGE in defining the fields of existence of the pseudo-loci and have extended the plausibility of these fields of existence by introducing fuzzy restrictions (softly bounded intervals) instead of binary restrictions (sharply bounded intervals).

The vector of breakpoints for F3P is:

V(F3P) = [2200, 2400, 2700, 3200, 3400]

These same fuzzy sets are defined for F3F. For F2P and F2S the fuzzy restrictions $\{l_{10}, a_{10}, b_{10}, c_{10}, d_{10}, e_{10}, f_{10}, h_{10}\}$ are defined with the following vector of break points:

V(F2P) = [700, 800, 1000, 1400, 1700, 1900, 2300, 2400]

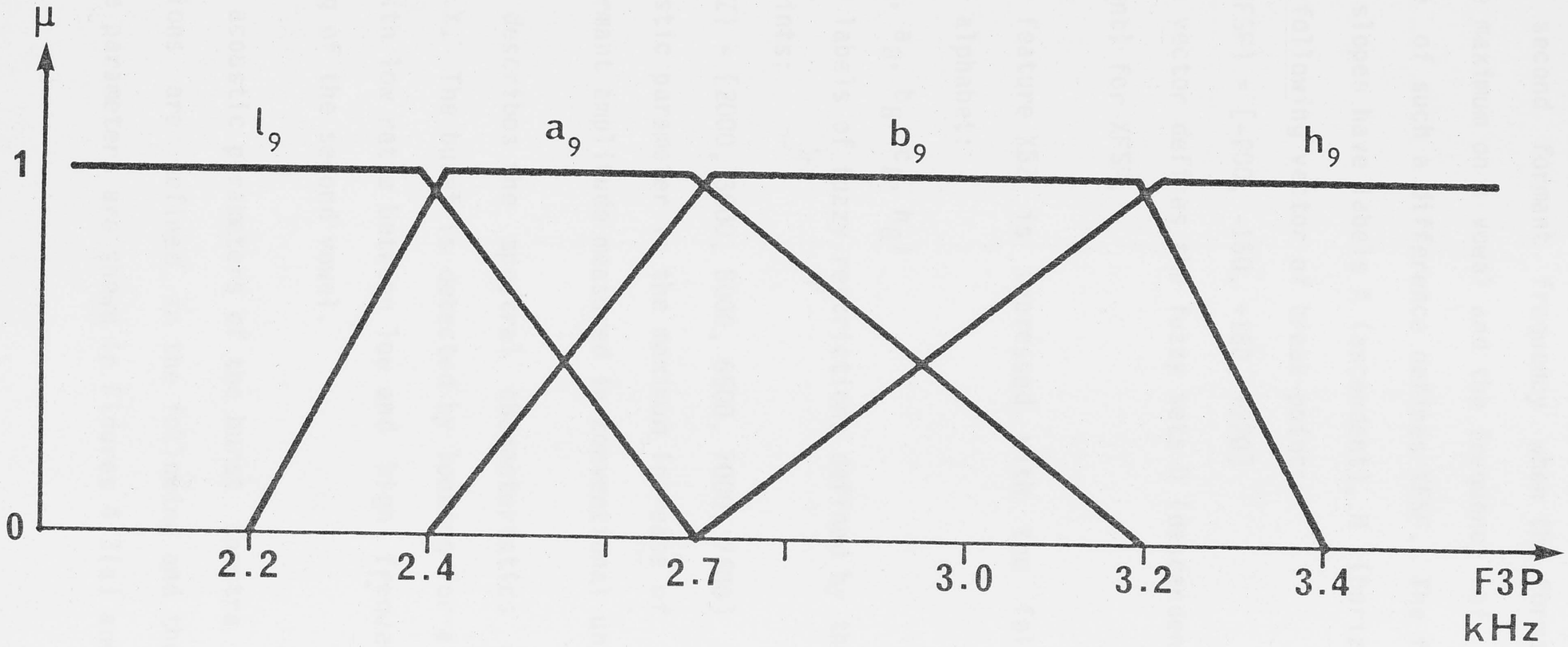


FIGURE 4.2: Fuzzy restrictions defined over F3P

The formant slope XFSF is defined as the difference between the value of the second formant frequency when the formant amplitude reaches the absolute maximum on a vowel and the frequency of the pseudo-locus; the opposite of such a difference defines XFSF. The fuzzy restrictions of the formant slopes have labels A (ascendent), H (horizontal), D (descendent) and the following vector of break-points:

V(XFSF) = [-200, -160, +160, +220]

The same vector defines the fuzzy sets D (descendent), H (horizontal), A (ascendent) for XFSP.

The feature XBZ is expressed with the following symbols of the terminal alphabet:

$\{l_8, a_8, b_8, c_8, h_8\}$

they are labels of fuzzy restrictions defined by the following vector of break-points:

V(BZ) = [2000, 3000, 5000, 6500, 7000, 7500]

the acoustic parameter is the maximum increase of the derivative of the first formant amplitude measured in conventional units.

XSP describes the spectral characteristics of the burst for the plosive X. The burst is detected by looking for a short peak in the total energy with low ratio between low and high frequency energy before the beginning of the second vowel.

The acoustic parameters of the burst spectra and the corresponding restrictions are defined in the following and the fields of the plosives for these parameters are shown in Figures 4.3(a) and 4.3(b).

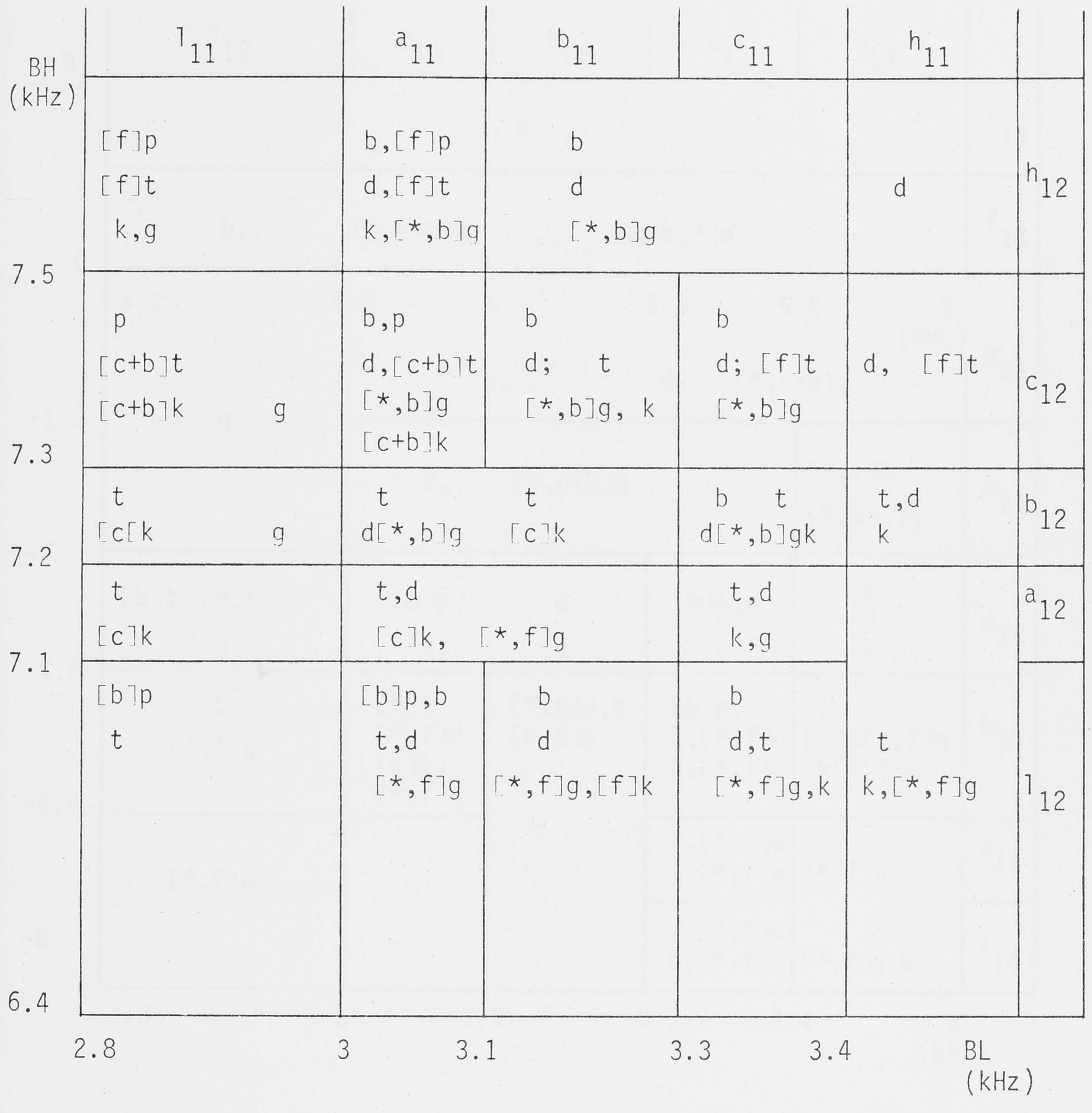


FIGURE 4.3(a): Fields of existence in the space defined by BH, the centre of gravity in the 5-10 kHz interval, and BL, the centre of gravity in the 1-5 kHz interval, of the bursts of plosive consonants occurring in the vocalic contexts indicated by the square brackets. Fuzzy interval labels are also marked.

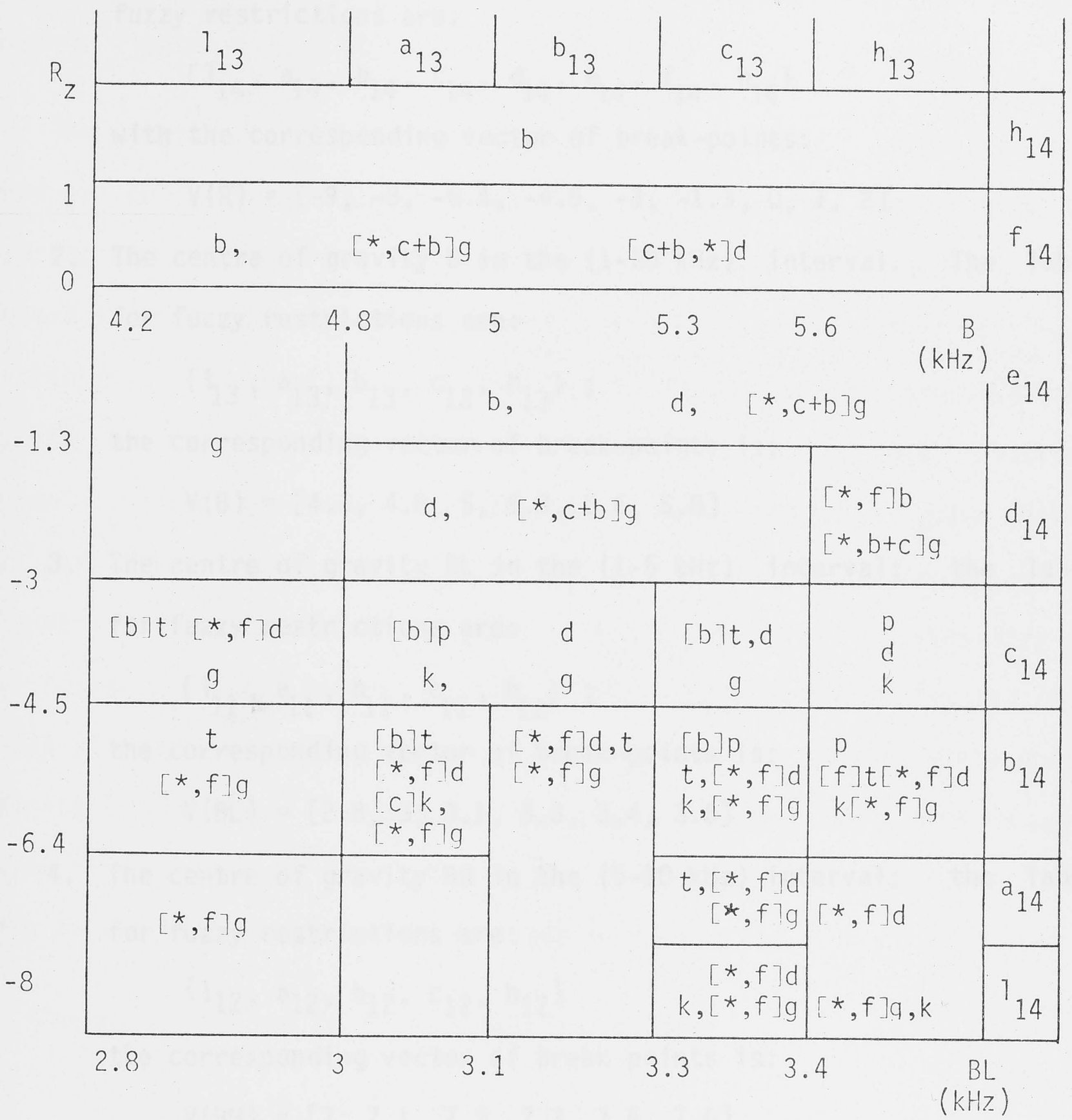


FIGURE 4.3(b): Fields of existence in the space defined by R, the ratio between low and high frequency energies, and B, the centre of gravity in the 1-10 kHz interval, of the bursts of plosive consonants occurring in the vocalic contexts indicated in the square brackets. Fuzzy interval labels are also marked.

1. Ratio R between low (299-900 Hz) and high (5-10 kHz) frequency energies measured in convenentional units. The labels for the fuzzy restrictions are:

$$\{l_{14}, a_{14}, b_{14}, c_{14}, d_{14}, e_{14}, f_{14}, h_{14}\}$$

with the corresponding vector of break-points:

$$V(R) = [-9, -8, -6.4, -4.5, -3, -1.3, 0, 1, 2]$$

2. The centre of gravity B in the (1-10 kHz) interval. The labels for fuzzy restrictions are:

$$\{l_{13}, a_{13}, b_{13}, c_{13}, h_{13}\} ;$$

the corresponding vector of break-points is:

$$V(B) = [4.6, 4.8, 5, 5.3, 5.6, 5.8]$$

3. The centre of gravity BL in the (1-5 kHz) interval; the labels for fuzzy restrictions are:

$$\{l_{11}, a_{11}, b_{11}, c_{11}, h_{11}\} ;$$

the corresponding vector of break-points is:

$$V(BL) = [2.8, 3, 3.1, 3.3, 3.4, 3.6]$$

4. The centre of gravity BH in the (5-10 kHz) interval; the labels for fuzzy restrictions are:

$$\{l_{12}, a_{12}, b_{12}, c_{12}, h_{12}\}$$

the corresponding vector of break points is:

$$V(BH) = [7, 7.1, 7.2, 7.3, 7.5, 7.6]$$

5. Finally, a set of restrictions is defined on consonant duration DURC; the labels are:

$$\{l_5, a_5, z_5, h_5\};$$

the corresponding break-points are the elements of the following vector:

$$V(DURC) = [0, 50, 70, 100, 150]$$

4.4.3 The Rules of the Fuzzy Grammar of Plosives

Most of the rules of the fuzzy grammar are derived from the diagrams shown in Figures 4.1(a)-(b) and 4.3(a)-(b).

The vocalic context is indicated inside brackets. For the tense plosives only the following vowel is considered; for the lax plosives, both the preceding and the following vowels are taken into account; the context is represented by the place of articulation of the vowel (b: back; c: central; f: front). For example [b,f] means that the preceding vowel has to be back and the following vowel has to be front. Places of articulation are defined as fuzzy sets on a two-dimensional space having the first two formants on the principal axes. The vocalic hypotheses are generated and evaluated after the detection of the zones of stability of the formants and the gravity centres of these zones are computed and their grades of membership with the fuzzy sets of the places of articulation are evaluated. The symbol + represents the logical disjunction and the * means 'every type of vowel'. The symbol is for negation. The rules are given using an algebraic notation for the sake of simplicity.

The rule relating to formant pseudo-loci are:

$$\begin{aligned}
\text{BFLP} = {} & [f,*](l_9+a_9)h_{10} + [f,f](l_9+a_9)(f_{10}+d_{10}) \\
& + [f,c](l_9+a_9)c_{10} + [c,f]l_9c_{10} + [f,b]a_9c_{10} \\
& + [b,f+c](l_9+b_9)b_{10} + [c,*]l_9b_{10} + [f,b]a_9b_{10} \\
& + [b,f+b](l_9+b_9)a_{10} + [b,b](l_9,b_9)l_{10}
\end{aligned}$$

$$\begin{aligned}
\text{DFLP} = {} & [f,f]a_9h_{10} + [f,*]a_9(f_{10}+d_{10}) + [c,*]a_9c_{10} \\
& + [b,*]a_9b_{10} + [b,f+b]a_9(a_{10}+l_{10})
\end{aligned}$$

$$\text{GFLP} = [f,*](a_9+b_9+h_9)(h_{10}+f_{10}) + [c,*]l_9(d_{10}+c_{10}) + [b,*](l_9+h_9)b_{10}$$

$$BFLF = [c,f]a_9f_{10} + [f,f](a_9+b_9+h_9)f_{10} + [f,f]d_{10} + [c+b,f]c_{10} + ([*,b] + [*,c] + [b,f])(a_{10}+b_{10}) + [*,b]l_{10}$$

$$DFLF = [*,f]f_{10}(a_9+b_9+h_9) + [c+f,f]d_{10} + (\neg b,b\ )c_{10} + [*,b](a_{10}+b_{10}) + [b,b]l_{10}$$

$$GFLF = [*,f](h_{10}+a_9f_{10}) + [f,c]d_{10} + [b,c]c_{10} + ([b,*] + [c,b])(a_{10}+b_{10}) + [b,*]l_{10}$$

Combining these rules, we arrive at the final formant pseudo-loci rules for the lax plosives:

$$BFL = BFLP.BFLF$$

$$DFL = DFLP.DFLF$$

$$GFL = GFLP.GFLF$$

The formant pseudo-loci rules for the tense plosives are:

$$PFL = [f]a_9f_{10} + ([c]+[b])l_9(a_{10}+b_{10}) + [b]l_9l_{10}$$

$$TFL = [f]f_{10}(a_9+b_9+h_9) + [f](l_9+a_9)d_{10} + [c](b_9+h_9)d_{10} + [f+c]c_{10} + [b](a_{10}+b_{10}+l_{10})$$

$$KFL = [f](h_{10}+(d_{10}+f_{10})(b_9+h_9)) + [c](l_9+a_9)d_{10} + [b](a_{10}+b_{10})$$

The buzz-bar rules are:

$$BBZ = [8,f](a_8+b_8+c_8+h_8)(z_5+d_5) + [*,c]a_5(b_8+c_8) + [*,b](a_8+b_8)(z_5+a_5)$$

$$DBZ = l_5((a_8+b_8) + [*,c]c_8 + [*,b](b_8+c_8+h_8))$$

$$GBZ = [*,f+c](l_8+a_8) + [*,b](l_8+a_8)$$

The formant slope rules are:

BFSP = [f+c,*]D + [b,*](H+D)

BFS = [*,f+c]A + [*,b](A+H)

BFS = BFSP.BFSF

DFSP = [b,*](A+H) + [c,*]H + [f,*](D+H)

DFSF = [*,b]D + [*,C]H + [*,f](A+H)

DFS = DFSP.DFSF

GFSP = [f,*]A + [b,*](H+D) + [c,f+c]A + [c,b](H+D)

GFSF = [*,f](H+D) + [*,c]D + [*,b](D+H)

GFS = GFSP.GFSF

PFS = [f+c]A + [b](A+H)

TFS = [b]D + [c]H + [f](A+H)

GFS = [b]D + [c]H + [f](A+H)

The rules for the burst spectra are the following:

$$BSP = h_{14} + 0.4f_{14} + (0.6e_{14}(a_{13}+b_{13}+c_{13}+h_{13}) + [*,f]d_{14}h_{13}) \cdot (a_{11}+b_{11}+c_{11})(h_{12}+c_{12}+b_{12}+l_{12}) + a_{12}(a_{11}+b_{11})$$

$$DSP = 0.4[c+b,*]f_{14} + ((a_{13}+b_{13}+c_{13}+h_{13})(0.6e_{14}+d_{14}+c_{14}) + [*,f]b_{14}) + [*,f](a_{14}.h_{13}+c_{13}.l_{14}))((a_{11}+b_{11}+c_{11}+h_{11}) \cdot (h_{12}+c_{12}+b_{12}) + (a_{11}+b_{11}+c_{11})(a_{12}+l_{12})$$

$$GSP = 0.4[c+b,*]f_{14} + (0.6e_{14}(l_{13}+[*,c+b](a_{13}+b_{13}+c_{13}+h_{13})) + [8,c+b]d_{14}(a_{13}+b_{13}+c_{13}+h_{13}) + c_{14}(l_{13}+a_{13}+b_{13}+c_{13}) + [*,f](b_{14}+(l_{13}+c_{13}+h_{13})a_{14}+c_{13}l_{14}))((l_{11}[*,b](a_{11}+b_{11}+c_{11})) \cdot (h_{12}+c_{12}+b_{12}) + ([*,f](a_{11}+b_{11})+c_{11})a_{12} + [f,*]l_{12} \cdot (a_{11}+b_{11}+c_{11}+h_{11}))$$

$$PSP = ((\lceil b \rceil(a_{13}+b_{13})+h_{13})c_{14} + (\lceil b \rceil c_{13}+h_{13})b_{14})$$
$$\cdot ((l_{11}+a_{11})(h_{12}\lceil f \rceil+c_{12})+l_{12}\lceil b \rceil(l_{11}+a_{11}))$$

$$TSP = (\lceil b \rceil(l_{13}+c_{13})c_{14}+(l_{13}+\lceil b \rceil a_{13}+b_{13}+c_{13}+\lceil f \rceil h_{13})b_{14}+c_{13}a_{14})$$
$$\cdot (\lceil f \rceil(l_{11}+a_{11})h_{12}+(\lceil c+b \rceil(l_{11}+a_{11})+b_{11}\lceil f \rceil(c_{11}+h_{11}\cdot c_{12}+b_{12})$$
$$+ (l_{11}+a_{11}+b_{11}+c_{11})a_{12}+(l_{11}+a_{11}+c_{11}+h_{11})l_{12})$$

$$KSP = ((a_{13}+b_{13}+h_{13})c_{14}+(\lceil c \rceil a_{13}+c_{13}+h_{13})b_{14}+h_{13}a_{14}+c_{13}l_{14})$$
$$\cdot (l_{11}+a_{11})h_{12} + (\lceil c+b \rceil(l_{11}+a_{11})+b_{11})c_{12}$$
$$+ (\lceil c+b \rceil l_{11}+a_{11})+b_{11})c_{12} + (\lceil c \rceil(l_{11}+a_{11}+b_{11}+c_{11}+h_{11})b_{12}$$
$$+ (\lceil c \rceil(l_{11}+a_{11}+b_{11})+c_{11})a_{12} + (\lceil f \rceil b_{11}+c_{11}+h_{11})l_{12})$$

A simple rule is very seldom useful for hypothesis generation is used alone. Rather, rules have to be combined in order to ensure that the right hypothesis is chosen in preference to the competing hypotheses.

An algebraic representation of the rules is given in the following. The grammaticalities have been dediced in order to give higher weight to the conjunction of robust features.

$$BFOR = 0.8.BFL + 0.85.BFL.BFS + 0.9\lceil c+b,c+b \rceil BFL.BBZ + BFL.BFS.BBZ$$

$$DFOR = 0.8.DFL + 0.85.DFL.DFS + 0.9\lceil c+b,c+b \rceil DFL.DBZ + DFL.DFS.DBZ$$

$$GFOR = 0.8.GFL + 0.85.GFL.GFS + 0.9\lceil c+b,c+b \rceil GFL.GBZ + GFL.GFS.GBZ$$

$$PFOR = 0.8.PFL + 0.95.PFL.PFS$$

$$TFOR = 0.8.TFL + 0.95.TFL.TFS$$

$$KFOR = 0.8.KFL + 0.95.KFL.KFS$$

Now the rules for generating hypotheses about phonemes are the following:

/b/ = 0.7BFOR + 0.6BSP + BFOR.BSP

/d/ = 0.7DFOR + 0.6DSP + DFOR.DSP

/g/ = 0.7GFOR + 0.6GSP + GFOR.GSP

/p/ = 0.7PFOR + 0.6PSP + PFOR.PSP

/t/ = 0.7TFOR + 0.6TSP + TFOR.TSP

/k/ = 0.7KFOR + 0.6KSP + KFOR.KSP

4.4.4 Example

The Figure 4.4 shows the formants of the syllable /adu/ extracted from continuous speech. The algorithms that will be described in following sections have given the following results:

$\mu_{<SONORANT>} = 0.7$

$\mu_{<TENSE>} = 0.6$

$\mu_{<NONSONORANT\ CONTINUANT>} = \mu_{<NONSONORANT\ AFFRICATE>} = 0$

$\mu_{<LAX>} = 0.8$

$\mu_{<NONSONORANT\ INTERRUPTED>} = 0.62$

$\mu_{<NONSONORANT>} = 0.85$

Thus, the fuzzy algorithm for the nonsonorant-interrupted-lax is applied first. The formants are tracked and the pseudo-loci (marked by circles in Figure 4.4) are extracted.

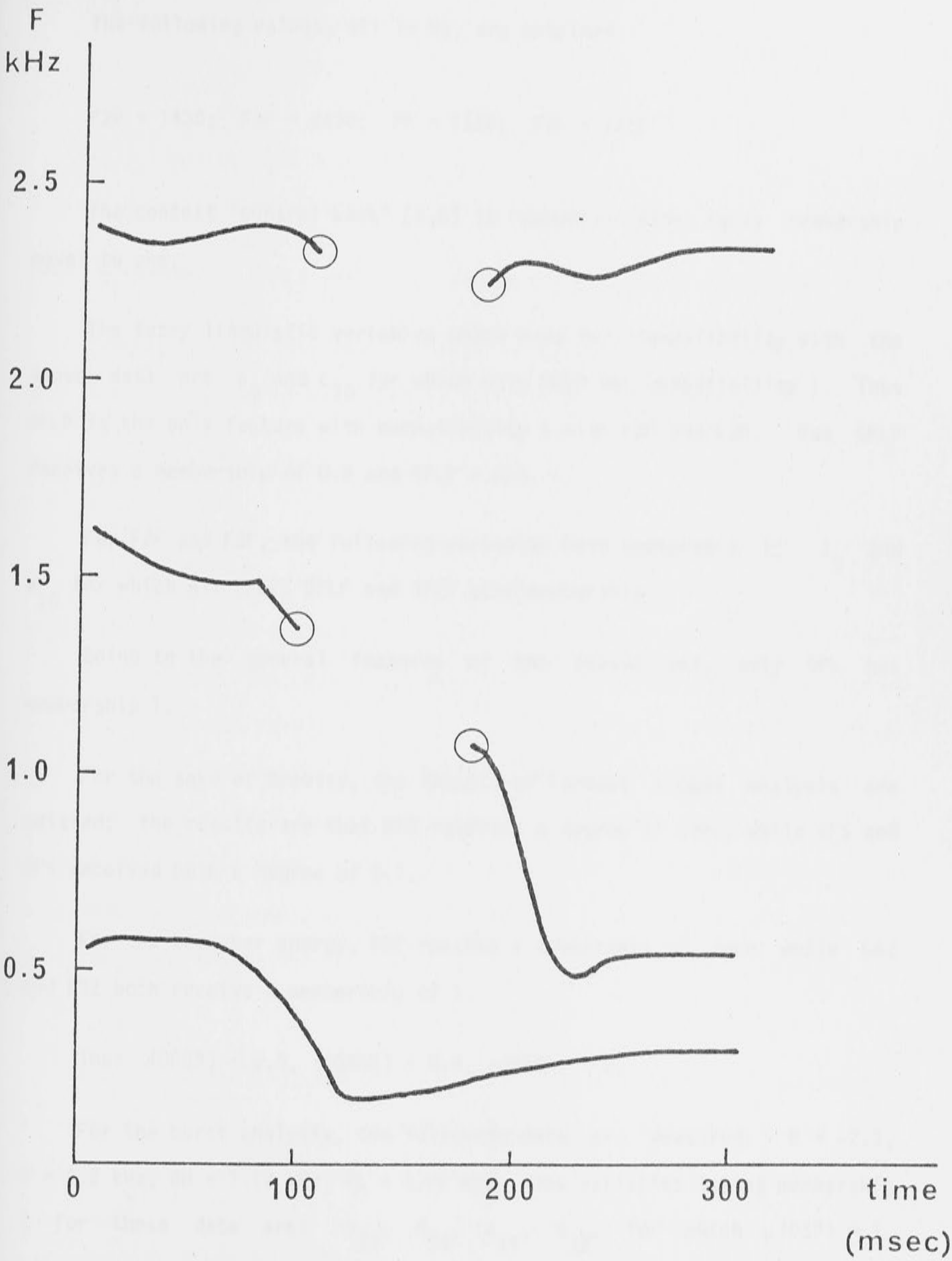


FIGURE 4.4: Formant transitions for the syllable /adu/.

The following values, all in Hz, are obtained:

F2P = 1430; F3P = 2430; 2F = 1110; F3F = 2350

The context 'central back' [c,b] is recognised with fuzzy membership equal to one.

The fuzzy linguistic variables which have full compatibility with the above data are a_9 and c_{10} for which only DFLP has compatibility 1. Thus DFLP is the only feature with compatibility 1 with F2P and F3P. But GFLP receives a membership of 0.9 and BFLP = 0.3.

For F2F and F3F, the following variables have membership 1; l_9 and b_{10} for which all BFLF, DFLF and GFLF have membership 1.

Going to the general features of the pseudo-loci, only DFL has membership 1.

For the sake of brevity, the details of formant slopes analysis are omitted; the results are that BFS receives a degree of zero, while GFS and DFS received both a degree of 0.7.

For the buzz-bar energy, BBZ reaches a membership of zero while GBZ and DBZ both receive a membership of 1.

Thus μ(DFOR) = 0.9, μ(GFOR) = 0.9, μ(BFOR) = 0.

For the burst analysis, the following data are measured: R = -2.3, B = 5.2 kHz, BH = 7.12 kHz, BL = 3.02 kHz. The variables taking membership 1 for these data are: b_{13}, d_{14}, a_{11}, a_{12}, for which μ(DSP) = 1, μ(BSP) = 0.3, μ(GSP) = 0.2.

From these data one gets:

$\mu(b) = 0.7\ 0.7\ 0.6\ 0.3\ 0.7\ 0.3 = 0.7$

$\mu(d) = 0.7\ 0.9\ 0.6\ 1\ 0.9\ 1 = 0.9$

$\mu(g) = 0.7\ 0.9\ 0.6\ 0.2\ 0.9\ 0.2 = 0.7$

and the only hypothesis accepted is (d).

4.5 DETECTION OF OTHER PHONETIC FEATURES OF NONSONORANT CONSONANTS

In this section the classification of consonants that occurs before the plosive recognition algorithm is applied is considered.

4.5.1 Precategorised Classification of Intervocalic Consonants

Most of the rules for the recognition of plosive sounds take account of the vocalic context. Although these rules have been inferred after experiments on vowel-consonant-vowel utterances extracted from continuous speech, most of the context dependencies refer to the previous or to the following vowel but rarely to both. This suggests that selected subsets of the rules can be used for plosives in contexts other than VCV selecting, for example, the rules depending on the previous vowel for contexts such as vowel-plosive consonant-?. As yet the possibility of extension has not been verified experimentally. Only the details referring to VCV contexts will be given here although investigations on their possible extension are in progress.

Precategorical classification of consonants consists of the assignment of phonetic features to nonvocalic segments. This arrangement is a generation of hypotheses using rules that do not require that the context be known. After the evidence for these hypothesis is evaluated, segmentation of continuous speech into pseudo-syllable segments (PSS) is performed (De Mori et al., 1976) and syllable bounds are used as edges for

a 'context-dependent' extraction of more detailed features used for phoneme hypothesization.

4.5.2 Hypothesization of the Features 'sonorant' and 'nonsonorant'

The classification of a consonant as sonorant or nonsonorant is performed after the detection of vocalic intervals (De Mori et al., 1976) and is obtained as an answer to a composite question.

Each question admits two possible answers: low or high. These answers will be indicated as: l_{1i}, h_{1i} where the subscripts 1i (i=1,2, ..., 6) refer to the answer of the 1i-th question: it is assumed that:

The answer to such questions are fuzzy restrictions defined over the ranges in which the parameters they refer to may vary. Such parameters are defined as follows:

- u_{11} = min R_v in the consonant
- u_{12} = min $(R_{vp} - R_{vc}; R_{vt} - R_{vc})$
- u_{13} = min $(S - S_{sil})$, $S \in S_{dip}$
- u_{14} = consonant duration
- u_{15} = min $(S_p - S_i; S_f - S_c)$
- u_{16} = max $(R_{vp} - R_{vc}; R_{vf} - R_{vc})$

where the subscript p refers to the detected vowel preceding the interval in which a nonsonorant feature is being sought; the subscript f refers to the detected vowel following the interval, the subscript c refers to the consonant interval and S_{sil} is the level of the r.m.s. signal in the silences.

The membership functions were defined after considering the range of the above parameters for each consonant in every context assigning values other than 1 to $\mu_{h_{1i}}$ and $\mu_{l_{1i}}$ (i=1,2,...,6) only in the range where sonorant and nonsonorant sounds may coexist. Notice that:

$$\mu_{l_{1i}} = 1 - \mu_{h_{1i}}$$

The answers to atomic questions are related to the values of hypotheses about phonetic features by fuzzy rewriting rules.

A set of fuzzy rules has been inferred for defining the syntactic categories sonorant and nonsonorant. The rules can be found in a paper by De Mori and Laface (1980) and are omitted here for the sake of brevity.

4.5.3 - Hypothesization of the Features 'continuant', 'interrupted' and 'affricate'

The features 'continuant', 'affricate', and 'interrupted' are further hypothesized for nonsonorant consonants. The motivations for this selection come from the evidence obtained by spectrogram reading experiments and from the possibility of deriving fast, simple and efficient algorithms for their hypothesization. For the same reasons, the possibility of hypothesizing two other features, namely, 'tense' and 'lax' for nonsonorant consonants has been introduced.

Although rules for the generation of hypotheses about these features have been proposed in the paper De Mori et al. (1976), a new set of rules is described here. These new rules are a refinement of the previous ones and have been obtained after a new round of experiments. As they are still 'context-independent' and are less abstract, showing most clearly what acoustic features are more strongly related to which phonetic features,

their details are given in the following.

The acoustic parameters utilised and the corresponding restrictions are defined as follows:

1. $S_{dip} - S_{sil}$: The level of the total energy in the dip with respect to the level of noise in the silences, measured in conventional units.

   The label for the fuzzy restrictions are:

   $$\{l_1, a_1, b_1, c_1, d_1, e_1, f_1, g_1, h_1\}$$

   and the corresponding VBP is:

   $$V(S_{dip}-S_{sil}) = [0,50,300,400,1000,1800,3000,4200,4500,5000]$$

2. R : The minimum of the ratio R in the dip, measured in conventional units.

   The labels for the fuzzy restrictions are:

   $$\{l_3, a_3, b_3, c_3, d_3, e_3, f_3, h_3\}$$

   The corresponding VBP is:

   $$V(R_{vmib}) = [-15000,-14000,-12000,-7000,-3000,-1000,0,1500,2000]$$

3. SR: The ratio between the minimum level of the total energy S in the dip plus a constant and the minimum value of R in its dip plus a constant measured in conventional units.

   The label for the fuzzy restrictions are:

   $$\{l_2, a_2, b_2, c_2, d_2, e_2, f_2, g_2, i_2, j_2, k_2, h_2\}$$

   The corresponding VBP is:

   $$V(SR) = [-2,-1,0,0.3,0.5,1,1.5,2,2.5,7,9,12,14]$$

4. $t_{R_V dip} - t_{Sdip}$: The delay of the dip in total energy S with respect to the dip in R_V, measured in msec.

   The labels for the fuzzy restrictions are:

   $$\{l_4, a_4, b_4, c_4, d_4, e_4, h_4\}$$

   The corresponding VBP is:

   $$V(t_{R_V dip} - T_{Sdip}) = [-150,-100,-30,0,10,60,100,120]$$

5. CDVR: The consonant duration measured in msec.

The labels for the fuzzy restrictions are:

$\{l_5, a_5, b_5, c_5, d_5, e_5, h_5\}$

The corresponding VBP is: V(CDUR) = [0,40,70,100,140,180,220]

6. R DUR: The duration of the dip in R measured in msec.

The labels for the fuzzy restrictions are:

$\{l_6, a_6, b_6, h_6\}$

The corresponding VBP is: $V(R_{vdip}DUR)$ = [0,2,5,8,10]

7. BBE: The buzz-bar energy measured in conventional units.

The labels for the fuzzy restrictions are:

$\{l_7, a_7, b_7, h_7\}$

The corresponding VBP is: V(BBE) = [-12000,-11000,-100000,-7000,12000]

The rules are given in the following using the algebraic notations of the preceding section.

<NONSONORANT INTERRUPTED>

= DDTE.(VVLTE.NGR2+NHTE.DSPDR.LR2)
+ HFE.(RTHR1+NVLR1.RTLTE.DSPDR.RTLDC)
+ DRFDS(l_6.DSNPDR)+0.6.DDTE.NHR2.NHTE

<NONSONORANT AFFRICATE>

= DDTE.(VVLTE.HR2.DSPDR+0.6.HR2.RTLTE)
+ DRFDS.(NAT_d+NAL_d) + HFE.(ADRDS).l_1

<NONSONORANT CONTINUANT>

= 0.6.DDTE.(h_1+h_4+a_4.b_5+LDURC.b_4) + HFE.(h_4+DRPDS)

<INTERRUPTED TENSE>

= 0.9.NBZ+0.9.BRT+NBZ.BRT+0.8.ULT.HTG

+ 0.8.ULB.AVTG+0.8ULB.LTG.LDURC+0.8HLFE.HTG.HDURC

+ 0.8.HLFE.LTG.HDURC+0.7RLLFE.HTG+0.7.RLLFE.AVTG

+ 0.7.RLLFE.UBR+0.6.RLLFE.VLTG+0.6HLFE.HTG.b_5

+ 0.6.HLF.UBR.c_1

<INTERRUPTED LAX>

= CBZ.BRL+0.9.CBZ+0.9.BRL+0.8.UBL.UBR

+ 0.7.UBL.LTG.VLDURC+0.7RLLFE.RLTG+0.6.HLFE.HTG.VLDURC

+ 0.7.HLFE.AVTG+0.6.HLF.UBR.RTHTE+0.6.HLFE.LTG.LDURC

<CONTINUANT TENSE>

= HFE.(DSPDRV+(LR_2.l_4|(f_1+h_5+l_6+(b_5.h_6)))

+ (a_4.(l_2+l_6+h_6)) + DDTE.(DRPDS+c_4)

<CONTINUANT LAX>

= HFE.((b_4+l_4)+(l_4+LR_2.(b_6+d_6)))

+ (l_4.a_2.AVCD.h_6) + (d_4.b_6) + DDTE.d_4

<AFFRICATE TENSE> = DDTE+NAL_d+HFE.MLLDC

<AFFRICATE LAX> = NAT_d+HFE.l_4.MLHDC

where:

NAT_d = DRFDS.(h_4+l_1+h_5+h_6)

NAL_d = DRFDS.(ADRDS.(LTE+l_6))

VLTE = l_1+a_1+b_1 : very low total energy

LTE = VLTE+c_1 : low total energy

RTLTE = LTE+d_1+e_1 : rather low total energy

NHTE = RTLTE+f_1 : not high total energy

HTE = f_1+g_1+i_1+h_1 : high total energy

RTHTE = d_1+e_1+HTE : rather high total energy

NLTE = RTHTE+c_1 : not low total energy

VLTE1 = a_1+b_1 : rather but not very low total energy

VVLTE = l_1+a_1 : very-very low total energy

DDTE : is a binary variable assuming value 1 when the dip in the total energy is deep (i.e. greater than 8000 conventional units).

LR2 = l_2+a_2+b_2+c_2 : low ratio

NHR_2 = LR2+d_2+e_2+f_2+g_2+i_2+j_2

: not high ratio

HR2 = k_2+h_2 : high ratio

RLR2 = HR2+i_2+j_2 : rather high ratio

MR2 = a_2+e_1 : medium ratio

AVR2 = MR2+f_2+g_2 : average ratio

VLR1 = l_3 : very low R_v minimum

NVLR1 = VLR1+a_3+b_3+c_3+d_3+e_3+f_3+h_3

: not very low R_v minimum

RTHR1 = d_3+e_3+f_3+h_3 : rather high R_v minimum

LR1 = VLR1+a_3 : low R_v minimum

HR1 = f_3+h_3 : high R_v minimum

VHR1 = h_3 : very high R_v minimum

HFE : is a binary variable assuming value 1 when minimum in R_v is low (i.e. less than -3000 conventional units).

DSPDRV = e_4+h_4 : the dip in total energy S precedes the dip in R_v very much

DSPRV = DSPDRV+c_4+d_4 : the dip in S precedes the dip in R_v

DSNPDR = $l_4+a_4+b_4+c_4$: the dip in S does not precede the dip in R_v

ADRDS = d_4+e_4 : average delay between dip in R_v and dip in S

DRPDS = $l_4+a_4+b_4$: dip in R_v precedes dip in S

DRFDS : is a binary variable assuming value 1 when the dip in R_v exists and it follows the dip in total energy S.

VLDURC = l_5+a_5 : very low consonant duration

LDURC = VLDURC+b_5 : low consonant duration

RTLDS = LDURC+c_5+d_5 : rather low consonant duration

MLLDC = $a_5+b_5+c_5+d_5$: more or less low consonant duration

AVCD = c_5+d_5 : average consonant duration

MLHDC = d_5+e_5 : more or less high consonant duration

HDURC = $d_5+e_5+h_5$: high consonant duration

RTHDC = HDURC+c_5+d_5 : rather high consonant duration

VLFFE = l_7 : very low buzz energy

LFFE = VLFFE+a_7 : low buzz energy

HLF = h_7 : high buzz energy

BRT = LR2.b_3 : tense-type burst

VLTG = $c_3 . l_3$: composite very low ratio

RLTG = RLR2.c_3 : composite rather low ratio

AVTG = AVR2.c_3 : composite average ratio

BRL = $b_2 . d_3 + f_2 . d_3$+RLR2.e_3+AVR2.e_3+RLR2.HR1+AVR2.HR1

: lax-type burst

RLLFE = $d_1 . a_7 + e_1 . a_7$+HTE.a_7

: composite rather low buzz energy

HLFE = HLF.l_1+HLF.c_1+RTHTE.HLF

: composite high buzz energy

ULB = VLTE.b_7+$c_1 . b_7$+$d_1 . b_7$

: composite burst

CBZ = 0.9.VLTE1.h_7+$e_1 . b_7$+HTE.b_7

: composite buzz energy

HTG = RHR2.c_3 : composite high ratio

UBR = $d_3 . l_2$+d_3.LR2 : uncertain burst

LTG = RLR2.c_3 : composite low ratio

NBZ = VLTE.LFFE+NLTE.VLFFE

: unevident buzz

4.5.4 Example

Figure 4.5 shows the time evolution of two parameters:

- the total energy S
- the ratio R between the values of energy in the low and the high frequency ranges for the same syllable /adu/ the formants of which are shown in Figure 4.4.

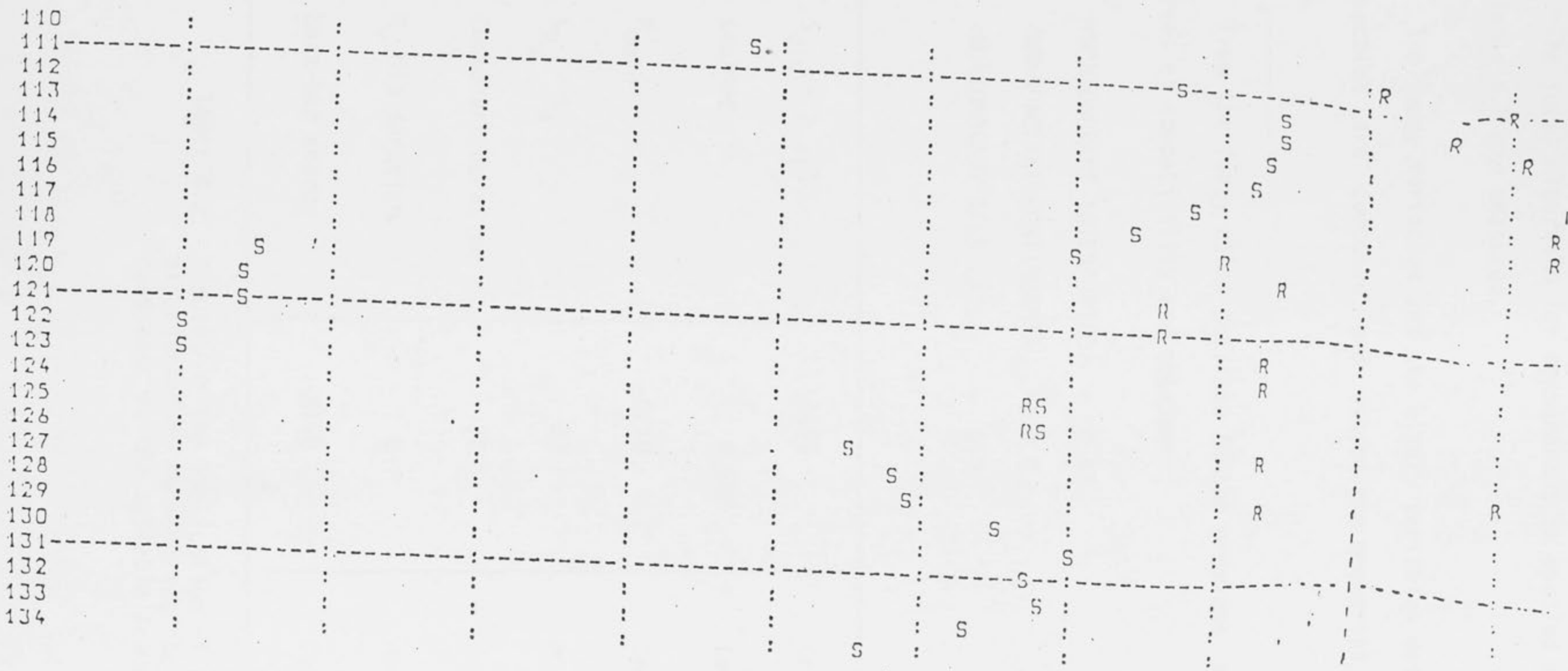


FIGURE 4.5: Evolution in time of the parameters S, the total energy, and R, the ratio between the low and high frequency energies, for the syllable /adu/.

The fuzzy algorithm for nonsonorant is applied and the values reported in Table 4.2 are measured.

The fuzzy variables and the binary variables defined in the preceding subsection are computed and assume the membership values shown in Table 4.3.

From the fuzzy rules outlined in the previous section the following degrees of compatibility are obtained:

<NONSONORANT INTERRUPTED> = 0.52

<NONSONORANT CONTINUANT> = 0.0

<NONSONORANT AFFRICATE> = 0.0

$S_{dip} - S_{sil}$	1245	Conventional units
$S+\alpha/R+\beta$	0.687	Conventional units
R_{vmin}	-2290	Conventional units
$t_{R_v} - t_s$	20	msec
Consonant duration	90	msec
R_v dip duration	0	msec
Buzz-bar energy	-8198	Conventional units

TABLE 4.2: Results of the application of the fuzzy algorithm for 'nonsonorant' to the consonant in the syllable /adu/.

L1	A1	B1	C1	D1	E1	F1	G1	H1				1
0.	0.	0.	0.68	1.	0.32	0.	0.	0.				
L2	A2	B2	C2	D2	E2	F2	G2	I2	J2	K2	H2	1
0.	0.	0.	0.63	1.	0.37	0.	0.	0.	0.	0.	0.	
L3	A3	B3	C3	D3	E3	F3	H3					1
0.	0.	0.	0.64	1.	0.36	0.	0.					
L4	A4	B4	C4	D4	E4	H4						1
0.	0.	0.	0.80	1.	0.20	0.						
L5	A5	B5	C5	D5	E5	H5						1
0.	0.33	1.	0.67	0.	0.	0.						
L6	A6	B6	H6									1
1.	0.	0.	0.									
L7	A7	B7	H7									1
0.	0.40	1.	0.60									
VLTE	LTE	RTLTE	NHTE	HTE	RTHTE	NLTE	VLTE1	VVLTE				1
0.	0.68	1.	1.	0.	1.	1.	0.	0.				
LR2	NHR2	HR2	RLR2	RHR2	MR2	AVR2						1
0.63	1.	0.	0.63	0.	1.	1.						
VLR1	NVLR1	RTHR1	LR1	HR1	VHR1							1
0.	1.	1.	0.	0.	0.							
DSPDRV	DSPRV	DSNPDR	ADRDS	DRPDS								1
0.20	1.	0.80	1.	0.								
VLDURC	LDURC	RTLDC	MLLDC	AVCD	MLHDC	HDURC	RTHDC					1
0.33	1.	1.	1.	0.67	0.	0.	0.67					
VLFFE	LFFE	HLF										1
0.	0.40	0.60										
BRT	VLTG	RLTG	AVTG	BRL	RLLFE	HLFE	ULB					1
0.	0.	0.63	0.64	0.36	0.40	0.60	1.					
CBZ	HTG	UBR	LTG	NBZ	NATA	NALA						1
0.32	0.	1.	0.63	0.	0.	0.						
DDTE	HFE	DRFDS										1
1	0	0										

TABLE 4.3: Membership values assumed by the fuzzy variables and binary variables used in plosive recognition when the algorithm is applied to the consonant in the syllable /adu/.

4.6 EXPERIMENTAL RESULTS AND CONCLUSIONS

The experimental results for the generation of phonetic and phonemic hypotheses depend on the scheduling policy adopted for the execution of the fuzzy algorithms.

Scheduling of a process is based on the evidence of the hypothesis which acted as a stimulus for the instantisation of a Knowledge Source (KS) resulting in that process.

Let U_1 be the universe of the acoustic parameters on which the hypothesis H_{ij} (j=1,2) has to be evaluated ($H_{11} \triangleq$ sonorant, $H_{12} \triangleq$ nonsonorant); let U_2 be the universe of the acoustic parameters on which the hypothesis H_{2j} (j=1,2,3) has to be evaluated ($H_{21} \triangleq$ continuant, $H_{22} \triangleq$ affricate, $H_{23} \triangleq$ interrupted); let U_3 be the universe of the acoustic parameters on which the hypothesis H_{3j} (j=1,2) have to be evaluated ($H_{31} \triangleq$ tense;, $H_{32} \triangleq$ lax); let U_4 be the universe of the acoustic parameters on which the hypotheses H_4 (j=1,2,...,6) have to be evaluated ($H_{41} \triangleq$ /p/; $H_{42} \triangleq$ /t/; $H_{43} \triangleq$ /k/; $H_{44} \triangleq$ /b/; $H_{45} \triangleq$ /d/; $H_{46} \triangleq$ /g/).

For the case of the plosive sounds, $U_3 > U_2 = U_1$; U_4 can be expressed by the following cartesian product:

$$U_4 = U_3 \times U_4'$$

and the rules introduced in Section 3 are defined over U_4' so as to help in differentiating the plosive sounds among themselves. In order to obtain absolute evidence of a plosive hypothesis H_{4j}, the following possibility has to be computed

Poss(H_{4j} is in p)

where $p \triangleq u \in U$ assuming that only the features in u characterise the pattern p when the possibility of H_{4j} has to be evaluated. Now we can write:

$u = u_3 . u_4'$ where

$u_3 \in U_3$, $u_4' \in U_4'$. So

(H_{4j} is in p) = (H_{4j} is in u_3) <and> (H_{4j} is in u_4').

Because of the mutual influence of the features in U_3 and U_4, the above two statements have to be considered as being INTERACTIVE. A reasonable rule for representing such an interaction is to express the total evidence as follows: Poss (H_{4j} is in p) = Poss (H^*_{3j} is in u_3).Poss(H_{4j} is in u_4').

where . is the arithmetic product; H^*_{3j} is the hypothesis consistent with H_{4j} (for example H_{4j} = lax if H_{4j} is /b/).

Poss H_{4j} is in u_4' is the possibility obtained by the algorithms described in Section 3.

Scheduling is based on the following rules:

Algorithms for H_{2j} are applied if:

$\mu_{NONSONORANT} - \mu_{SONORANT} > 0.3$

Algorithms for H_{3j} are applied if;

$\mu_{INTERRUPTED} - MAX\{\mu_{CONTINUANT}, \mu_{AFFRICATE}\} > 0.6$

Algorithms for P_L (lax plosives) are applied if;

$\mu_{LAX} - \mu_{TENSE} > -0.6$

Algorithms for P_T (tense plosives) are applied if;

$\mu_{TENSE} - \mu_{LAX} > 0.1$

Using these types of scheduling rules, the probability of missing the right hypothesis is less than 0.01, the average number of competing hypotheses among the possible nine is the following set $\mathcal{T}$ defined in the

following is 2.4 and the probability of having the right hypothesis with the highest membership is 0.86.

T = VOCALIC, SONORANT-LIQUID, SONORANT-NASAL,
NONSONORANT-INTERRUPTED-TENSE, NONSONORANT-INTERRUPTED-LAX,
NONSNORANT-CONTINUANT-TENSE, NONSONORANT-CONTINUANT-LAX,
NONSONORANT-AFFRICATE-TENSE, NONSNORANT-AFFRICATE-LAX#

For evaluating the performances of the algorithm for the plosive sounds, a set of about 500 syllabic muclei extracted from continuous speech was analysed. The nuclei were uttered by four male and one female speakers.

Using the simple technique of taking only one hypothesis if there is only one phoneme whose compatibility with the pattern is almost 0.1 higher than the memberships of the competing hypotheses and keeping all the hypotheses whose membership is higher than the membership of the most evident hypothesis minus 0.1, the following results have been achieved:

absence of the right hypothesis: less than 1%
generation of a single hypothesis: 80% of the cases
generation of two hypotheses: 15% of the cases
generation of three hypotheses: 3% of the cases.

Thus the average number of plosive hypotheses is 1.25. As the average number of hypotheses generated during the assignment of phonetic features is 2.4, the average number of phoneme hypotheses that would be produced when a plosive consonant is pronounced is about 3. This number should be augmented to take into account ambiguities in segmentation, hypothesization of consonant clusters and detailed feature extraction such as formant tracking. A factor of 2 or 3 is a good estimation of such an augmentation.

A new system for generating and verifying hypotheses about phonemes has been presented. It integrates all the designers' phonetic knowledge in its fuzzy rules and takes into account the vagueness of such a knowledge.

The imprecision of the data measured on the pattern is accounted for by describing the input data by fuzzy linguistic variables. The fuzziness of these variables models speaker differences fairly well.

The great redundancy of the features used for recognition seems to compensate the vagueness with which they are detected and the knowledge which controls their usage. Further investigations have to be carried out for inferring the knowledge of the structure of plosives in consonant clusters. The grammaticalities of the rules are expected to be refined during the execution of further experiments.

Chapter 5

AUTOMATIC RECOGNITION OF AUSTRALIAN ENGLISH VOWELS IN CONTINUOUS SPEECH

In this chapter the author's algorithm for automatic recognition of vowels in Australian English is presented. Particular attention is paid to the problems of inter-speaker differences. There is also some discussion of the need (or otherwise) of very precise recognition of vowels in continuous speech.

5.1 INTRODUCTION

The folklore about automatic recognition of vowels is that as a class vowels are more speaker dependent and less context dependent than any other class of speech sounds. Because of this, almost all vowel recognition algorithms rely on speaker-normalized data and do not make allowances for coarticulation effects. In recognition systems that use coarticulation effects to obtain good recognition scores the vowels provide 'anchor points', i.e. once the vowel has been recognized coarticulatory rules can be used to identify neighbouring sounds. Speaker normalization is generally achieved by having a speaker who is new to the system read a standard list of words and from this reading parameters are measured which can be fed to a vowel normalizing routine such as Gerstman's (1968). In this chapter the need for such normalization is considered in detail and it is concluded that while some normalization is needed, a requirement that the system be trained to each new speaker can be avoided. After all, the speech perceptual processes of the brain can understand without special

adaptation what a new speaker is saying provided that he speaks a language dialect with which the listener is familiar.

Another aspect of vowel recognition examined here is the effect of context. As was noted in Chapter 1, the perception of vowels in context is much more categorical than the perception of vowels in isolation (Stevens, 1968). We investigate the acoustic correlates of this effect.

Here too, the first of the algorithms for automatic recognition of Australian English speech is presented. Although a few linguistic studies on Australian English speech exist, there is no other work on the automatic recognition of this dialect. Vowel recognition algorithms developed for other varieties of English cannot be directly applied to Australian speech as it is the vowels of a dialect which most distinguish it from other dialects of the same language. That there are considerable differences between Australian English and British and American English (for which most of the vowel recognition algorithms have been developed to date) has been shown in several studies (Bernard, 1967a, 1970; Burgess, 1968; Hanley and Andrews, 1967; Wagner, 1978).

5.2 AUSTRALIAN VOWEL DATA USED IN THIS STUDY

Counting both monophthongs and diphthongs (but ignoring for the moment the not unknown triphthongs) it is generally agreed that there are twenty vocalic nuclei in Australian English. To obtain an adequate number of continuous speech samples for all of these nuclei in a variety of environments and for a large number of speakers would be a substantial task and one which was luckily unnecessary as there are sufficient studies of Australian English vowels available from which data can be pooled.

In his 1967a study of Australian English, Bernard recorded 170 male adults saying 19 vowel nuclei in the /h-d/ frame, once when the words occurred in isolation and once when they occurred in a stressed sentence-final position. Bernard analysed his data spectrographically and durationally. In another study carried out concurrently with Bernard's study, Burgess (1968) made spectrographic measurements on eight Australian English vowels occurring in a /b-t/ frame. Twenty male speakers read a piece of prose which contained at least one occurrence of the eight vowel sounds in the specified context. Recently two further sets of Australian vowel data have been collected. Oasa (1980) recorded lists of words which contained nearly all Australian vowels in a variety of contexts. He did spectrographic measurements on words containing the vowels /æ,a,ɒ,ɔ, and u/ as read by eighteen male and twelve female speakers. The other collection was made by the present author as part of an experiment described in detail in the next chapter. Ten speakers, five male and five female, produced at least 48 instances of the front vowel /i/ and the back vowel /ɔ/ in conversational-style speech. Each instance of the vowel occurred in a controlled phonetic environment. The vowels were extracted from conversational speech and analysed using a 12-coefficient for male voices and 16-coefficient for female voices linear prediction program. Results from the four studies are considered here. Each study provided a useful cross-check on the others and it was satisfying to note that all four gave comparable mean values for the common vowels investigated*. There were also other useful cross-checks. For example, possible shifts in vowel nuclei positions over the decade between the Bernard and Burgess studies on the one hand and the Oasa and the author's studies on the other were

* Or where differences were noticed they could be accounted for satisfactorily.

considered. Although there was some evidence for possible vowel shifts, this was not strong and did not significantly impede the reliability of pooling the results from the different studies when it was advantageous to do so. Details of the various studies are given in Table 5.1 and in Figure 5.1.

The four studies also complemented each other neatly. The Bernard, Burgess, and Oasa studies provided many clear instances of a large number of nuclei in citation-form words where the subject was producing the vowels in stressed (i.e. presumably non-reduced) positions. From these studies canonical positions of the Australian vowel nuclei can be measured. The author's data, however, provide examples of the way in which vowel parameters can deviate from their canonical forms in conversational speech with differing degrees of stress and with differing phonetic contexts. The data for the female speakers in the Oasa and the author's studies were useful in evaluating the problem of normalization between speakers needed for automatic recognition.

Nothing has yet been said of the famous Mitchell and Delbridge study on Australian English (1965). This study was actually very useful in that it provided the first systematic account of Australian English speech. However, as the data in this study were largely descriptive and non-numerical they could not be directly compared with the data cited in the studies described above, although it is freely acknowledged that the Mitchell and Delbridge study provided the impetus for the spectrographic studies.

DATA COLLECTED BY	VOWELS STUDIED	NUMBER AND SEX OF SPEAKERS	TYPE OF MATERIAL USED
Bernard (1969)	/i/, /ɪ/, /ɛ/, /æ/, /a/, /ɒ/, /ʊ/, /u/, /ʌ/, /ɔ/, /ɜ/, /eɪ/, /aɪ/, /ɔɪ/, /aʊ/, /oʊ/, /ɪʊ/, /ɛə/, /ʊə/	170 male	Vowels in /h-d/ frame. Each /h-d/ word was produced once in isolation and once in a sentence-final stressed position.
Burgess (1968)	/ɪ/, /ɛ/, /æ/, /a/, /ʌ/, /ɒ/, /ɔ/, /ʊ/	20 male	Vowels in /b-t/ frame. Each /b-t/ word was contained in a prose passage read by each subject.
Oasa (1980)	/æ/, /a/, /ɔ/, /u/	18 male 2 female	Several different CVC words were read from a list by each subject.
O'Kane	/i/, /ɔ/	5 male 5 female	Each subject was presented with a list of 48 two-word combinations and instructed to make up sentences containing these two-word combinations.

TABLE 5.1: Details of studies on Australian English vowels

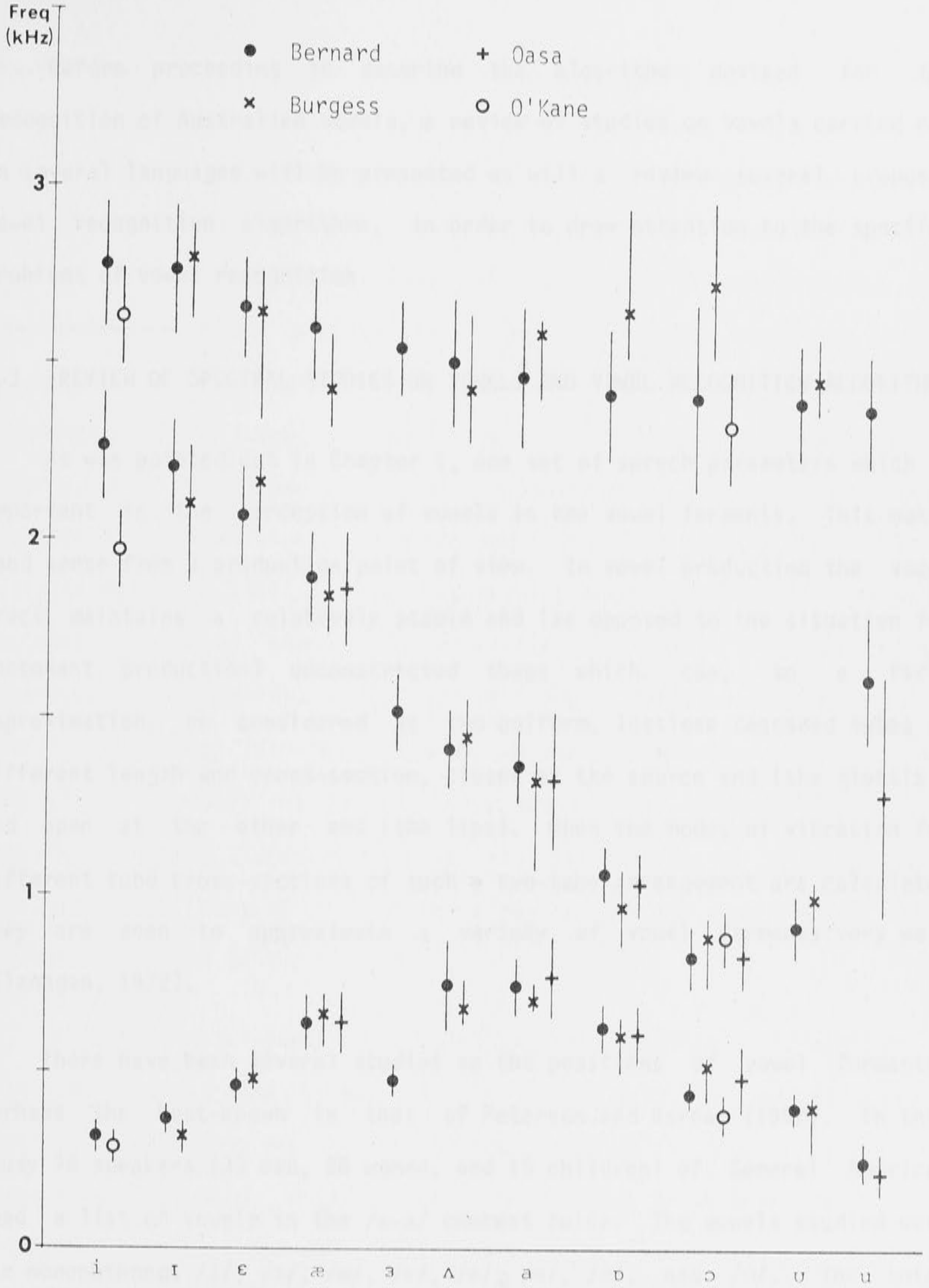


FIGURE 5.1: Mean and first standard deviation positions of F1, F2, and F3 of the vowels studied by Bernard (1967) and O'Kane, and of F1 and F2 of the vowels studied by Oasa (1980). Mean and range positions of F1, F2, and F3 of the vowels studied by Burgess (1968).
Note: Data in all cases averaged across all speakers in each data set.

Before proceeding to describe the algorithm devised for the recognition of Australian vowels, a review of studies on vowels carried out in several languages will be presented as will a review several proposed vowel recognition algorithms, in order to draw attention to the specific problems of vowel recognition.

5.3 REVIEW OF SPECTRAL STUDIES ON VOWELS AND VOWEL RECOGNITION ALGORITHMS

As was pointed out in Chapter 1, one set of speech parameters which is important in the perception of vowels is the vowel formants. This makes good sense from a production point of view. In vowel production the vocal tract maintains a relatively stable and (as opposed to the situation for consonant production) unconstricted shape which can, to a first approximation, be considered as two uniform, lossless cascaded tubes of different length and cross-section, closed at the source end (the glottis), and open at the other end (the lips). When the modes of vibration for different tube cross-sections of such a two-tube arrangement are calculated they are seen to approximate a variety of vowel formants very well (Flanagan, 1972).

There have been several studies on the positions of vowel formants. Perhaps the best-known is that of Peterson and Barney (1952). In this study 76 speakers (33 men, 28 women, and 15 children) of General American read a list of vowels in the /h-d/ context twice. The vowels studied were the monophthongs /i/, /ɪ/, /æ/, /ɜ/, /ʌ/, /ʊ/, /ɔ/, and /u/. The total 1520 words were all examined on the sound spectrograph and the values of the first three formants recorded. Several interesting results came from this study. It was found that when the results for all speakers were plotted in the F1-F2 plane (i.e. first formant versus second formant) the

vowels could be reasonably well separated although there was some overlap between the regions associated with the different vowels. When, however, data collected from several repetitions of the vowels by just one speaker were plotted in the F1-F2 plane, the vowels were well separated with no overlap. From this it was deduced that vowel formant positions were speaker-dependent with each speaker having an internally consistent method of overall vowel production. On the subject of speaker dependencies it was also found that children's formants are highest in frequency, women's intermediate, and men's lowest in frequency. As well as taking spectrographic measurements, Peterson and Barney also carried out listening tests in which the 1520 /h-d/ words were played to a panel of 70 listeners. The words were not unanimously identified as those intended by the speakers, the proportion of 'correct' identifications depending on the nature of the vowel intended. Thus 94% of the /i/ sounds were correctly identified but only 6% of the /ɒ/ sounds were correctly identified. Nevertheless, when listeners disagreed with speakers on the identity of a vowel the two classifications were almost always for vowels which occupy adjacent positions in the F1-F2 plane.

The experimental procedure adopted by Bernard (1967a) is similar to that of Peterson and Barney (1952). In Figure 5.1 the mean and one standard deviation points for the Australian English vowels studied by Bernard are plotted. Also plotted on this figure are similar results for the Australian English vowels studied by Burgess (1968), Oasa (1980), and the author.

Other studies on formant positions have been carried out. Pols, van der Kamp, and Plomp (1970) examined formant positions for 12 Dutch vowels occurring in a /h-d/ context. The aim of their study was to discover which

of the six parameters F1, F2, F3, L1, L2, L3 (where LX is the amplitude level of FX) were most effective in separating the different classes of vowels. They found that the list of parameters in order of usefulness in vowel classification is: log F2, log F1, log F3, L3, L2, L1. In other respects the results of the Pols et al. study confirmed the results of the Peterson and Barney study.

The usefulness of formant positions in obtaining good vowel classifications has been exploited in vowel recognition algorithms for automatic speech recognition. In a vowel recognition algorithm designed by Forgie and Forgie (1959) for the recognition of vowels in a /b-t/ frame, the positions of the first two formants were roughly located and then a table was consulted to see to which set of vowels the unknown set of vowels could belong. After this, more detailed measurements were made on various spectral features (mostly area under 'valleys' in the spectrum and slopes of formant peaks, pitch determination, and sometimes F3) and these were fed to special subroutines which separated out the vowels in the list of possibles to which the unknown had been assigned. The overall score for this system when tested on 21 subjects (11 male and 10 female) producing vowels in the /b-t/ context was 88%. This score was raised to 93% when some duration information was included.

In the acoustic-phonetic analysis system devised by Weinstein et al. (1975) vowels were located in running speech by measurements of the energy in various frequency bands. Steady-state vowels were then identified by comparing the formant positions at the centre of the vowel to a speaker-normalized formant table containing target formant positions. This system correctly identified vowels 41% of the time and correctly placed the unknown vowel in its top three choices 69% of the time. In this

system as in other ARPA systems, vowel recognition was always dependent on pre-training of the machine to a new speaker.

5.4 SPEAKER NORMALIZATION - HOW NECESSARY IS IT?

5.4.1 On-going system adaptation to new speakers

Here we query whether or not speaker normalization is really necessary. Certainly when procedures such as Gerstman's (1968) are used good recognition hand resulted. Gerstman, working with the Peterson as Barney data, scaled the data for the first and second formants of each speaker such that the lowest and the highest values were set equal to 0 and 999 respectively, and all other frequencies were then rescaled as linear extents between the two extremes. Then using an algorithm which used the parameters F1*, F2*, (F1+F2)*, and (F1-F2)* (where the * refers to the rescaled parameters) he was able to achieve 97.5% accuracy classification of vowels as the vowel intended by the speaker. No wonder Weinstein et al. (1975) used the Gerstman algorithm in their vowel identification procedure!

An interesting aspect of Gerstman's result is that his classification achieved better results than did the panel of listeners in the original Peterson and Barney experiments. It would seem that in the everyday multi-speaker environment listeners quite often do not hear the sound intended by the speaker. This probably would not matter too much because most of what is heard is heard in sentence context, thus higher level information can be used to determine the final identity of the vowel. And as the Peterson and Barney (1952) listening experiment showed, wrong vowel identifications are almost always 'near misses' so that if a listener has reason at some processsing stage to doubt his earlier vowel identification

he probably tries neighbouring vowels instead (where 'neighbouring' can refer to neighbouring along a variety of dimensions).

Yet there is evidence that some sort of vowel normalization does take place. Ladefoged and Broadbent (1957) found that when they played a carrier phrase 'Please say what this word is' followed by the test word, the identity of the vowel in the test word depended on the positions of the formants of the vowels in the carrier phrase and to a lesser extent on the linguistic background of the listener.

Taking together the various results discussed above, it seems that what is happening is that there are enough overall cues of a general nature (both phonetic and higher level) for a listener to obtain a fair idea of what is being said by a speaker who is new to him, but as the listener becomes more familiar with the speaker's voice he (the listener) is able to identify that speaker's vowels with increasing accuracy. In an automatic recognition system this approach could be implemented in the following way. An utterance from a new speaker would be analyzed according to general rules for phonetic recognition. However at the same time statistics about the voice could be gathered (often from an analysis of mistakes) and these statistics could be used to conduct a search through the files that the system holds on speakers that it has encountered before. If a match is found then the contents of that file will be used in conjunction with the general rules of phonetic analysis for further recognition of what that speaker says. If no match is found then the system will create a new file into which it will put information about this new speaker. When the system is faced with that particular speaker at some later time it can immediately retrieve the file for that speaker if he declares his identity, or proceed as for a new speaker if he does not.

In everyday speech communication we are constantly encountering the same people and we find that we know their voices very well. Similarly the automatic system will collect a great deal of information about familiar speakers perhaps even to the extent of developing certain special recognition algorithms for each of those speakers. The files for speakers that the system encounters infrequently will be stored in a low priority access area so that it will take some time before the system 'remembers' that it has encountered the speaker before. Also files that are in the low priority access area probably have been 'trimmed', that is only minimal information is kept in them and new information is collected when (and if) the speaker associated with the file is encountered again. There are a couple of further points to note on this topic in the light of the Foreign Phonetician model introduced in Chapter 1. It is commonly said that all foreigners from the same general area sound alike. In other words unless we know that the speaker is speaking a language/dialect with which we are familiar we do not extract information about the speaker per se as easily as if we were familiar with his dialect. This makes sense as it indicates that until we know the important parameters of a language we cannot estimate what is legitimate speaker variation within the confines of these parameters. Yet we seem capable of extracting some global parameters about the speaker which help in making decisions as to the geographical origins of that speaker. In a system which acts like a foreigner in a land where the natives speak their own language, the mechanism for extracting information about new speakers will not be as efficient as the native's extraction of such parameters although as the system attunes better and better to the language it is attempting to recognize, then the particular parameters which are relevant to speaker characterization will become clearer. An extensive discussion of which parameters these might be is

given by Wagner (1978).

We have discussed the means whereby an automatic recognition system could improve its performance in vowel recognition by on-going adaptation to a new speaker. We now go on to show that it is possible to design an automatic vowel recognition algorithm which requires no pre-training of the speaker to the system.

5.4.2 Normalization according to the sex of the speaker

Leaving aside the subject of adaptation to a particular speaker, let us consider how much normalization is necessary for vowel recognition. In Figure 5.2 and 5.3 we have given a display of the frequencies of the positions of the first three formants of the steady-state positions of the vowels /i/ and /ɔ/ for ten speakers, five male and five female. (As was mentioned earlier the experimental conditions under which this data was collected are described in detail in the next chapter.) For each speaker there are forty-eight instances of each vowel corresponding to forty-eight different but controlled phonetic and pausal environments. It should be noted that all the vowels are extracted from extemporaneous, conversational-style speech. Several things are immediately noticeable about these displays, in particular the following:

(1) The differences between male and female formants for these two vowels are most noticeable in the region above 1700 Hz. Thus the male/female distinction is very noticeable in the F2 and higher formants of /i/ and to a lesser extent in F3 and higher formants of /ɔ/.

(2) Within the sex groupings there is still a fair degree of variation. For example the second and higher formants of /i/ for Speaker 2 are in general considerably lower than those for the other male speakers

FIGURE 5.2: Mean and first standard deviation points of the first, second and third formants of the vowel /i/ for Speakers 1-10.

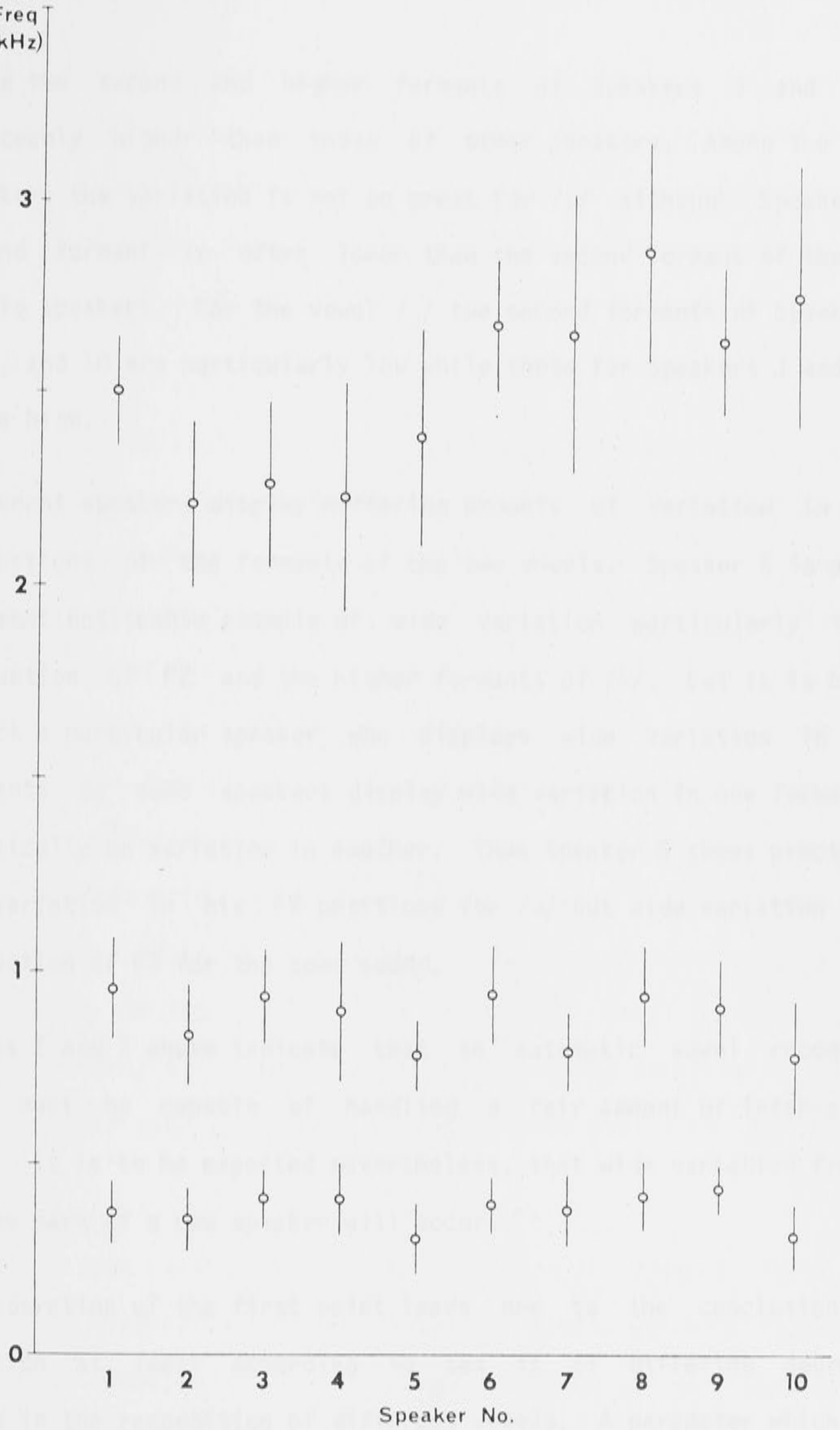


FIGURE 5.3: Mean and first standard deviation points of the first, second and third formants of the vowel /ɔ/ Speakers 1-10.

while the second and higher formants of Speakers 3 and 4 are noticeably higher than those of other speakers. Among the female speakers the variation is not so great for /i/ although Speaker 6's second formant is often lower than the second formant of the other female speakers. For the vowel /ɔ/ the second formants of Speakers 2, 5, 7, and 10 are particularly low while those for Speakers 3 and 6 are quite high.

(3) Different speakers display differing amounts of variation in their productions of the formants of the two vowels. Speaker 6 is perhaps the most noticeable example of wide variation particularly in her production of F2 and the higher formants of /i/. But it is hard to select a particular speaker who displays wide variation in vowel formants as some speakers display wide variation in one formant and practically no variation in another. Thus Speaker 5 shows practically no variation in his F2 positions for /ɔ/ but wide variation in his production of F3 for the same sound.

Points 2 and 3 above indicate that an automatic vowel recognition algorithm must be capable of handling a fair amount of inter-speaker variation. It is to be expected nevertheless, that wide variation from the norm on the part of a new speaker will occur.

Consideration of the first point leads one to the conclusion that normalization at least according to sex is of differing degrees of importance in the recognition of different vowels. A parameter which could be used to help in such normalization is pitch. Pitch should generally separate speakers according to sex.

Further clarification of point (1) above was obtained by considering sex differences in the vowels in Oasa's data in conjunction with the /i/-/ɔ/ data already considered. These data are illustrated in Figure 5.4. In this figure mean positions of the first and second formants of the vowels /i/, /æ/, /a/, /ɒ/, and /ɔ/ are shown for both male and female speakers.

Thus for vowels with high first formants (> 600 Hz at least), there is a strong likelihood that there will be frequency differences according to sex. A similar situation holds for the second formant although the trend is much more pronounced than in the case of the first formant. The higher the average value of the second formant of a vowel, the greater the difference between the average value of that formant for male and female speakers. A check through the averages for the Peterson and Barney (1952) data confirms this trend also for American vowels. Fant (1966) produced a physiological explanation for differences between male and female vowel formants on the basis of data he had collected from Swedish subjects, in conjunction with the Peterson and Barney (1952) data. Fant's (1966), Peterson and Barney's (1952), Oasa's (1980), and the author's data for the difference between male and female second formant positions is given in Figure 5.5.

One thing that Fant did not notice but which is evident form Figure 5.5 is that the difference between the average female and male formants for vowels is a non-linear monotonic increasing function of the average male second formant for vowels. Thus we see that there is no significant difference between the average positions of the male and female second formants for /ɔ/ which has a low second formant, but there is quite a

FIGURE 5.4: Average first and second formants of the vowels /i/, /æ/, /a/, /ɒ/, /ɔ/ for male and female voices.

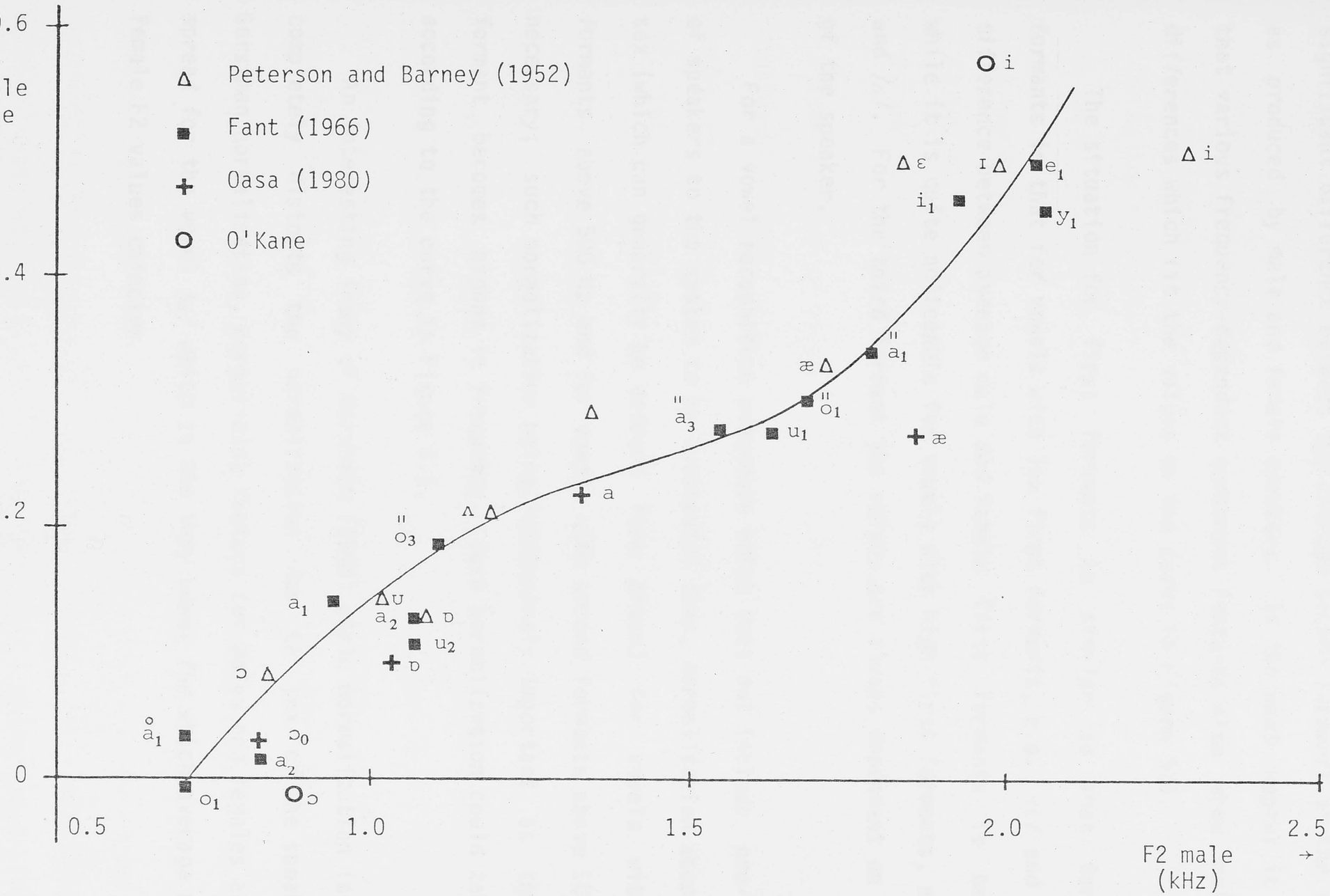


FIGURE 5.5: Difference between average female and male second formants for several vowels as a function of the average male second formant for those vowels

Note: Fant's (1966) symbols are used for his data. IPA symbols used in all other cases.

significant difference between the average second formants of the vowel /i/ as produced by male and female speakers. In the next chapter it is shown that various frequency-dependent consonant features also show male/female differences which fit the values on the curve in Figure 5.5.

The situation for first formants is similar to that for second formants in that for vowels with low first formants, e.g. /i/ and /ɔ/, the difference between average male and female first formants is negligible while it is quite noticeable for vowels with high first formants, e.g. /a/ and /ʌ/. For the third formant the values are always dependent on the sex of the speaker.

For a vowel recognition procedure which does not include pre-training of speakers to the system to be successful then, normalization according to sex (which can generally be deduced from pitch) for vowels with first formants above 500 Hz and for vowels with second formants above 1000 Hz is necessary; such normalization being increasingly important as the vowel formant becomes higher in frequency. Such normalization could take place according to the curve in Figure 5.5.

An interesting irony of Gerstman (1968) style normalization is that it completely distorts the normalization due to sex of the speaker. In Gerstman normalization, normalizing factors for males and females are most spread for the vowel /ɔ/ which is the very vowel for which average male and female F2 values coincide.

5.5 RELATIONSHIPS BETWEEN FORMANTS

In Figures 5.6 (a)-(c) and 5.7 (a)-(c), F1 versus F2, F2 versus F3, and F3 versus F4, for the vowels /i/ and /ɔ/ as spoken by the ten speakers in the author's data base are displayed. From these displays it is clear that there exist correlations between the following:

F1 and F2 for /ɔ/ - a weak correlation

F2 and F3 for /i/

F3 and F4 for both /i/ and /ɔ/

From a survey of Bernard's data (1967a) it was found that similar relationships hold between the second and third formants of all the front vowels, and the first and second formants of all the back vowels except /u/. This is illustrated in Figures 5.8 and 5.9.

Such relationships are useful as a consistency check on early hypotheses in automatic vowel recognition. Thus, for example, the positions of the formants of an unknown vowel might fall within the frequency range of each of those formants of the vowel /i/. Yet unless the F2 and F3 values of the unknown occur in the correct position in F3-F2 space for /i/, the vowel will not be classed as a potential /i/.

Figure 5.10 is a diagrammatic representation of fuzzy set defined for /i/ in F3-F2 space. The membership of this fuzzy set is 1 within the inner ellipse; is between 0 and 1 between the two ellipises (the actual value depending on the distance from the inner ellipse), and 0 outside the outer ellipse. High membership in this fuzzy set will ensure that (for male voices) an unknown vowel has fulfilled the consistency condition for /i/.

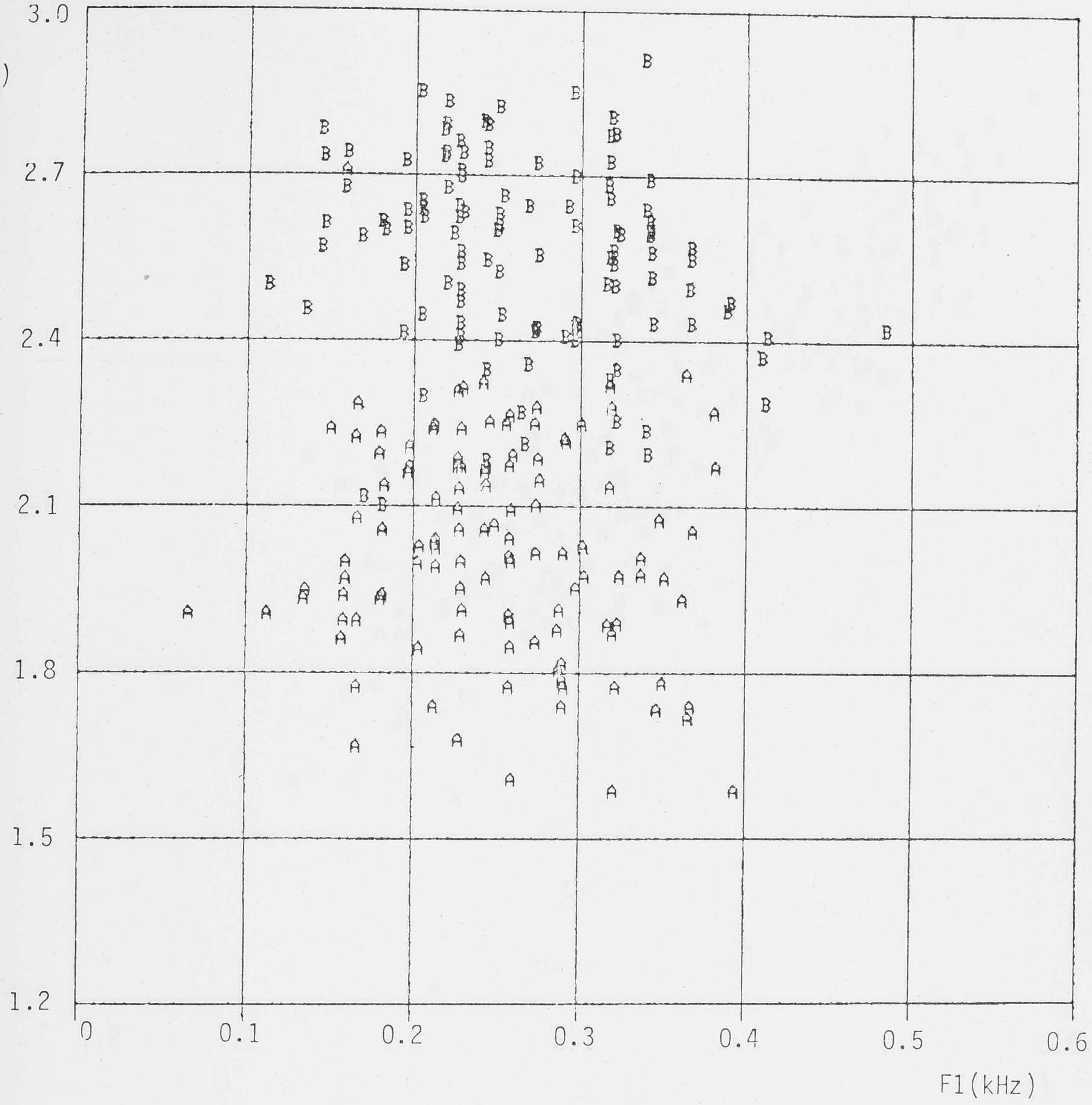


FIGURE 5.6(a): F2 versus F1 for the vowel /i/.

Note: For this and all the other diagrams of Figures 5.6 and 5.7, the following applies:

A: male voices
B: female voices

There were 5 male and 5 female speakers. There are 24 examples of each speaker's production of each vowel.

Problems with formant measurement add variance to the data. For example, a first formant frequency below 150 Hz should be regarded as suspect (since the vocal-tract walls place a lower limit in F1).

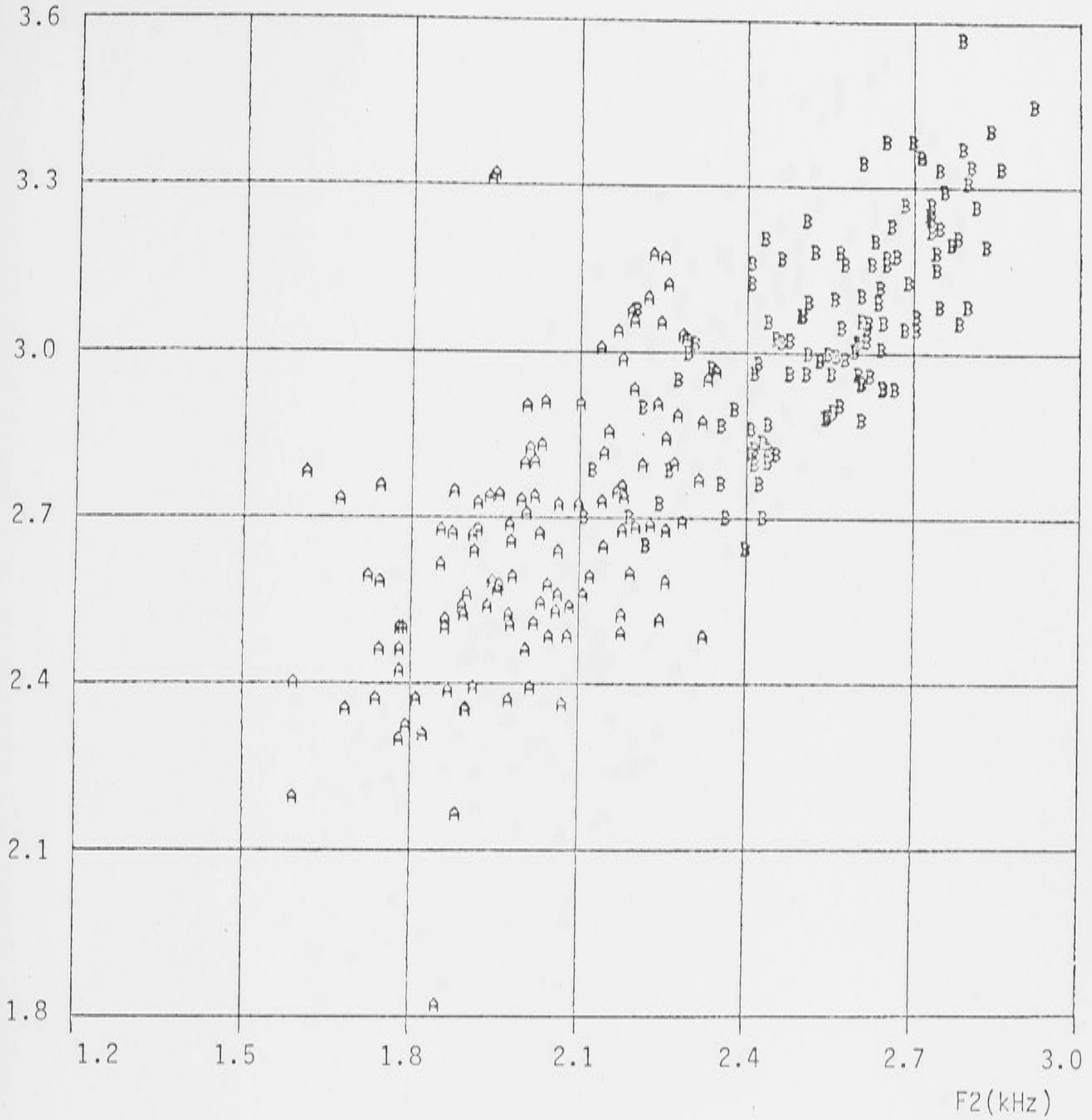


FIGURE 5.6(b): F3 versus F2 for the vowel /i/.

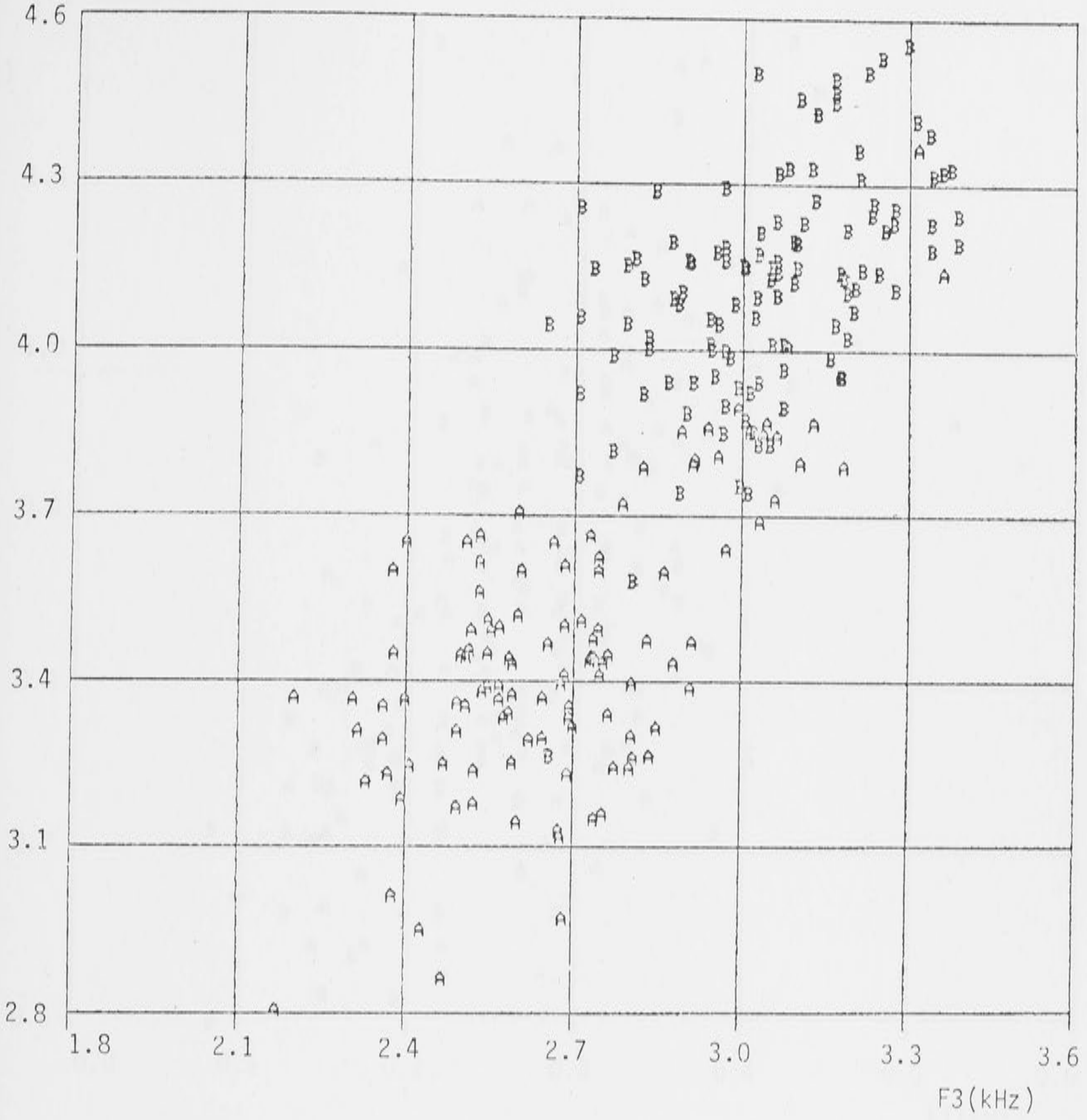


FIGURE 5.6(c): F4 versus F3 for the vowel /i/.

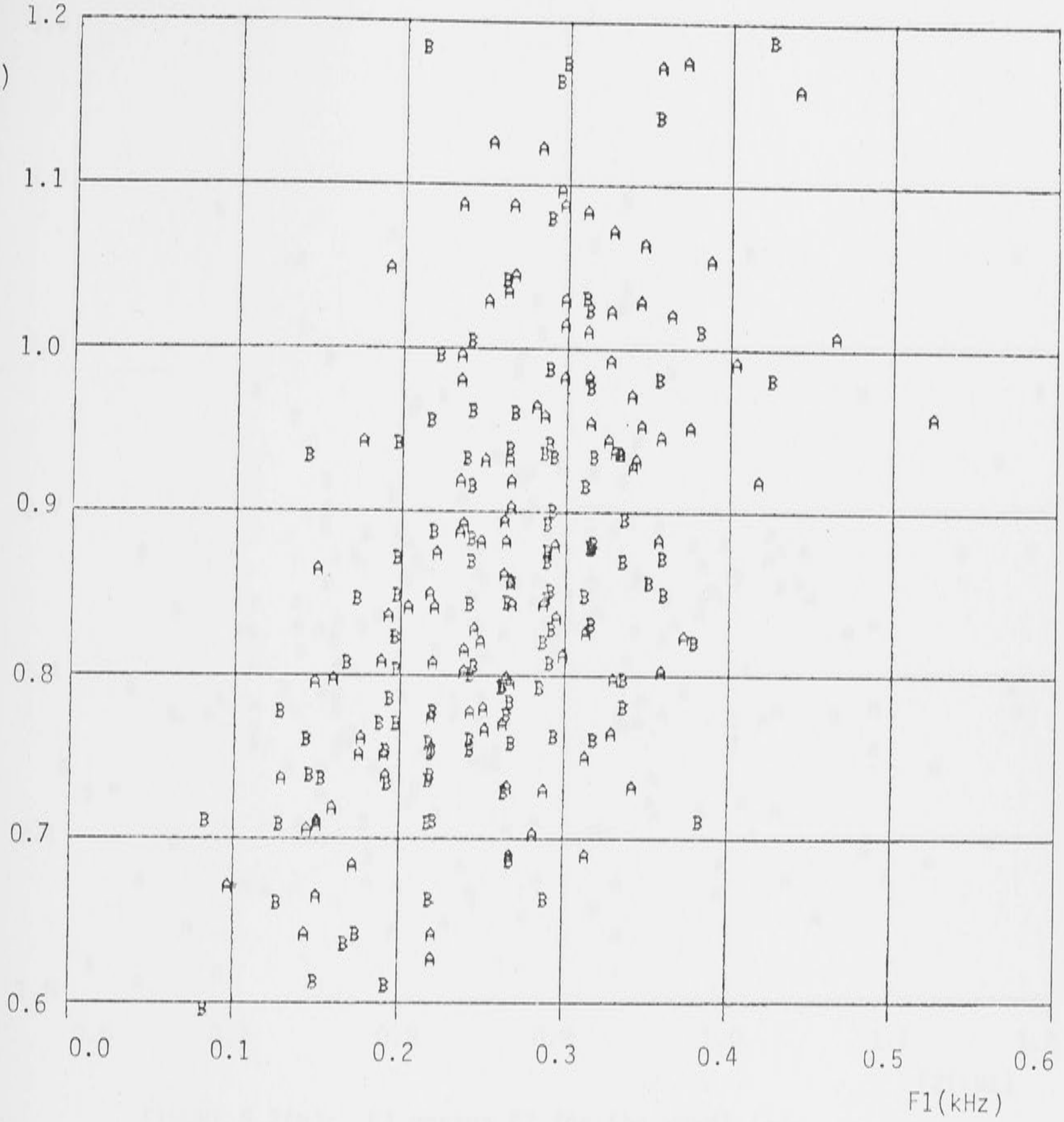


FIGURE 5.7(a): F2 versus F1 for the vowel /ɔ/.

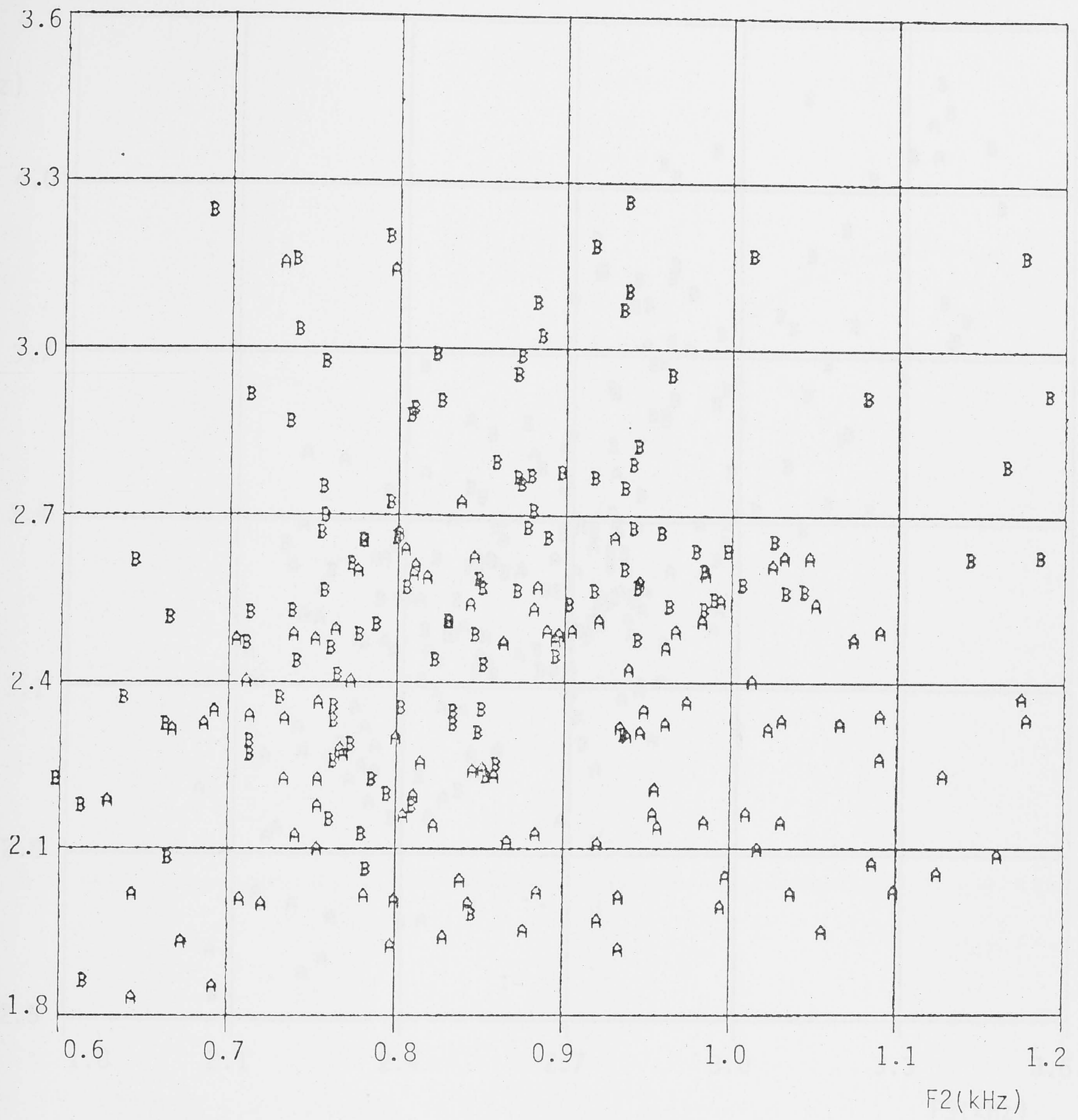


FIGURE 5.7(b): F3 versus F2 for the vowel /ɔ/.

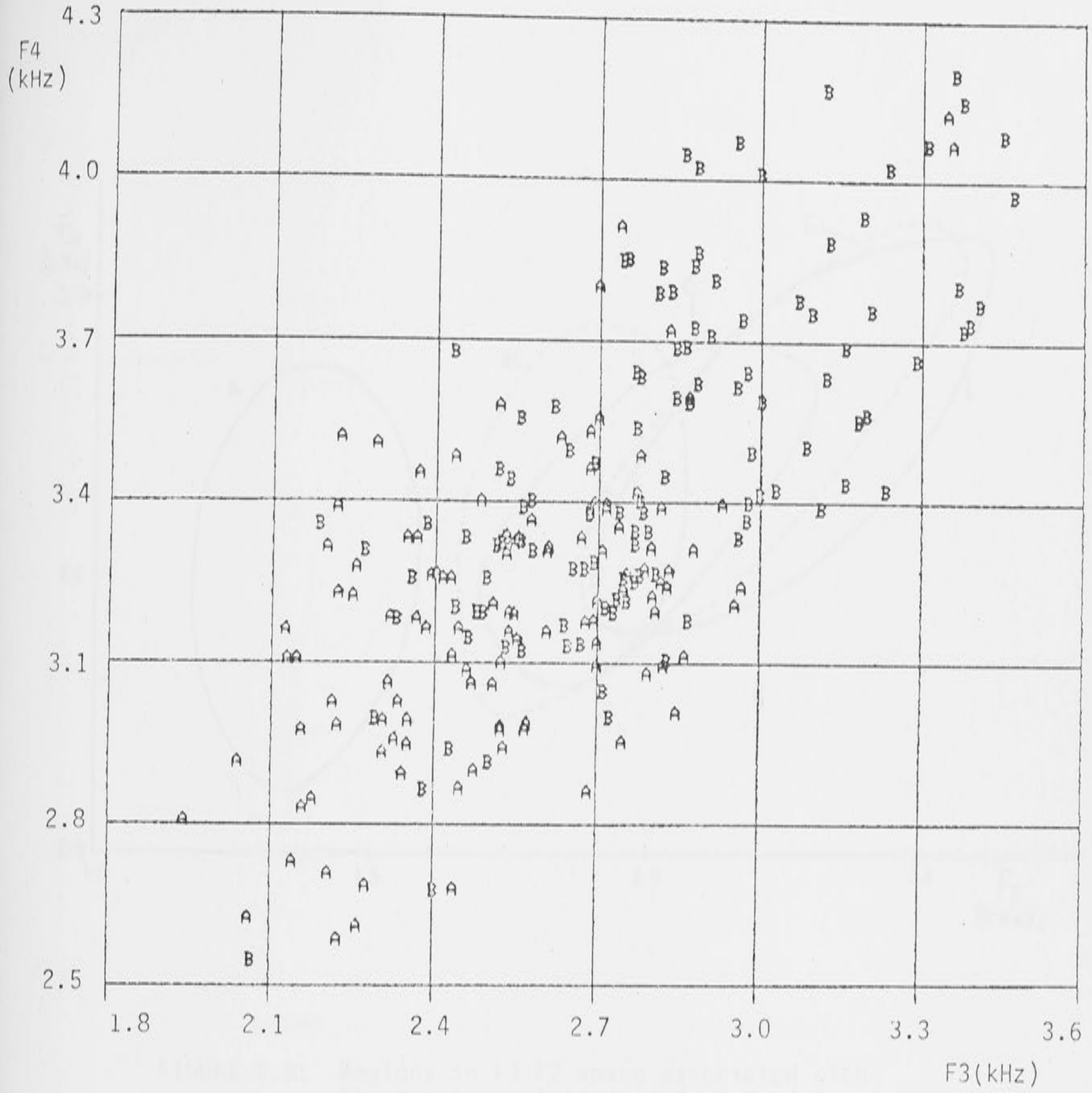


FIGURE 5.7(c): F4 versus F3 for the vowel /ɔ/.

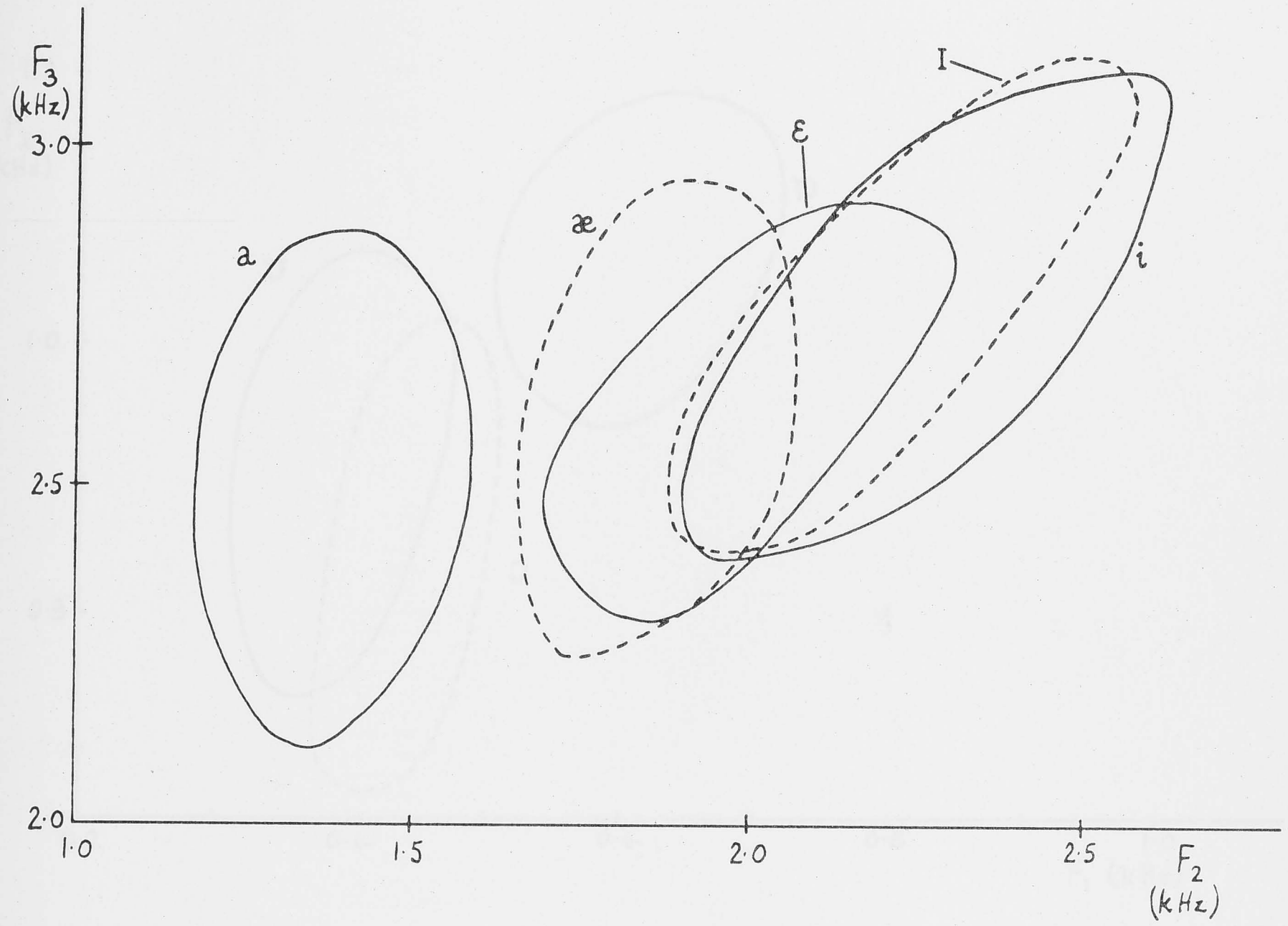


FIGURE 5.8: Regions in F3-F2 space associated with the front vowels /i/, /ɪ/, /ɛ/, /æ/, /a/.

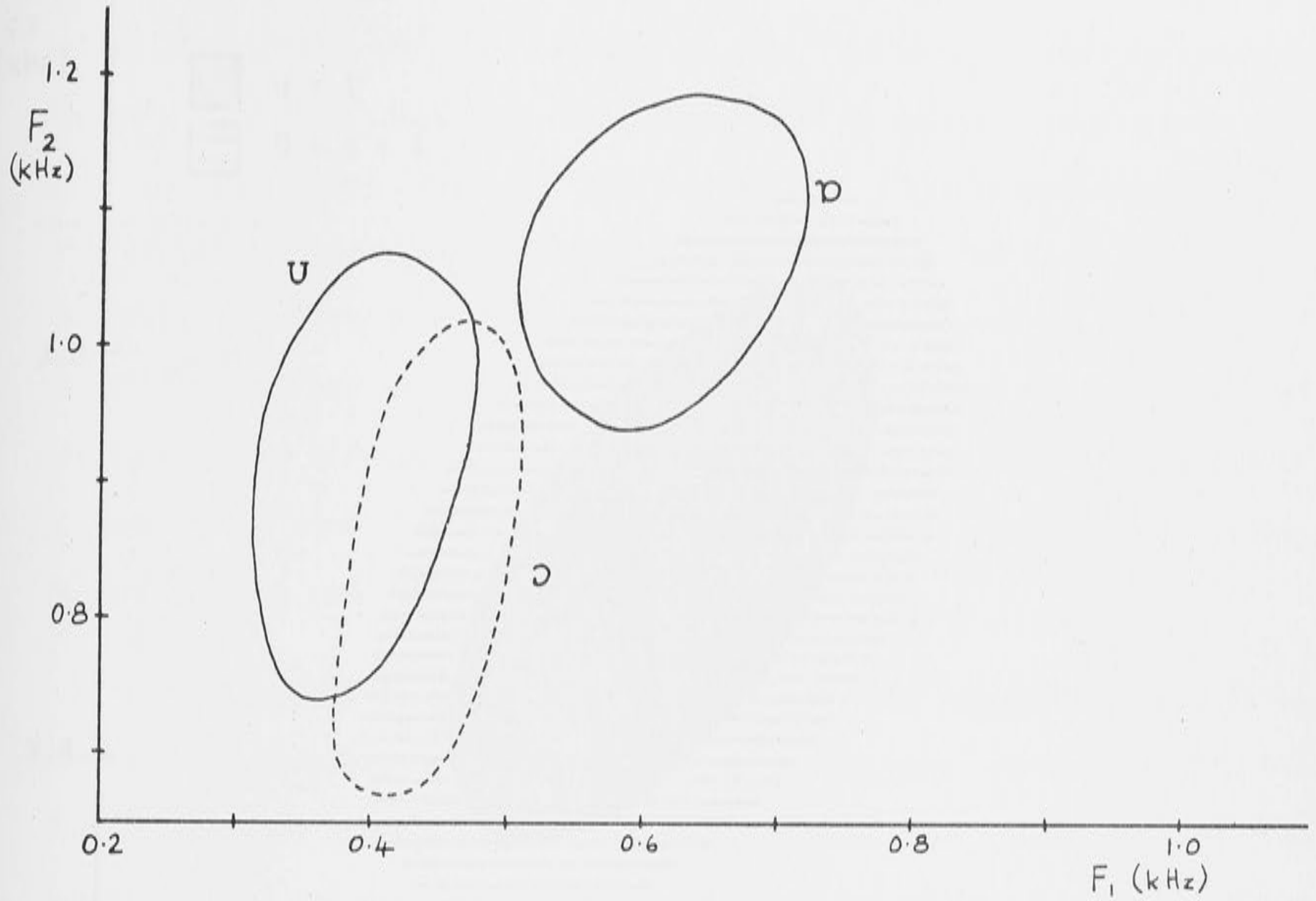


FIGURE 5.9: Regions in F2-F1 space associated with the back vowels /ɒ/, /ʊ/, /ɔ/.

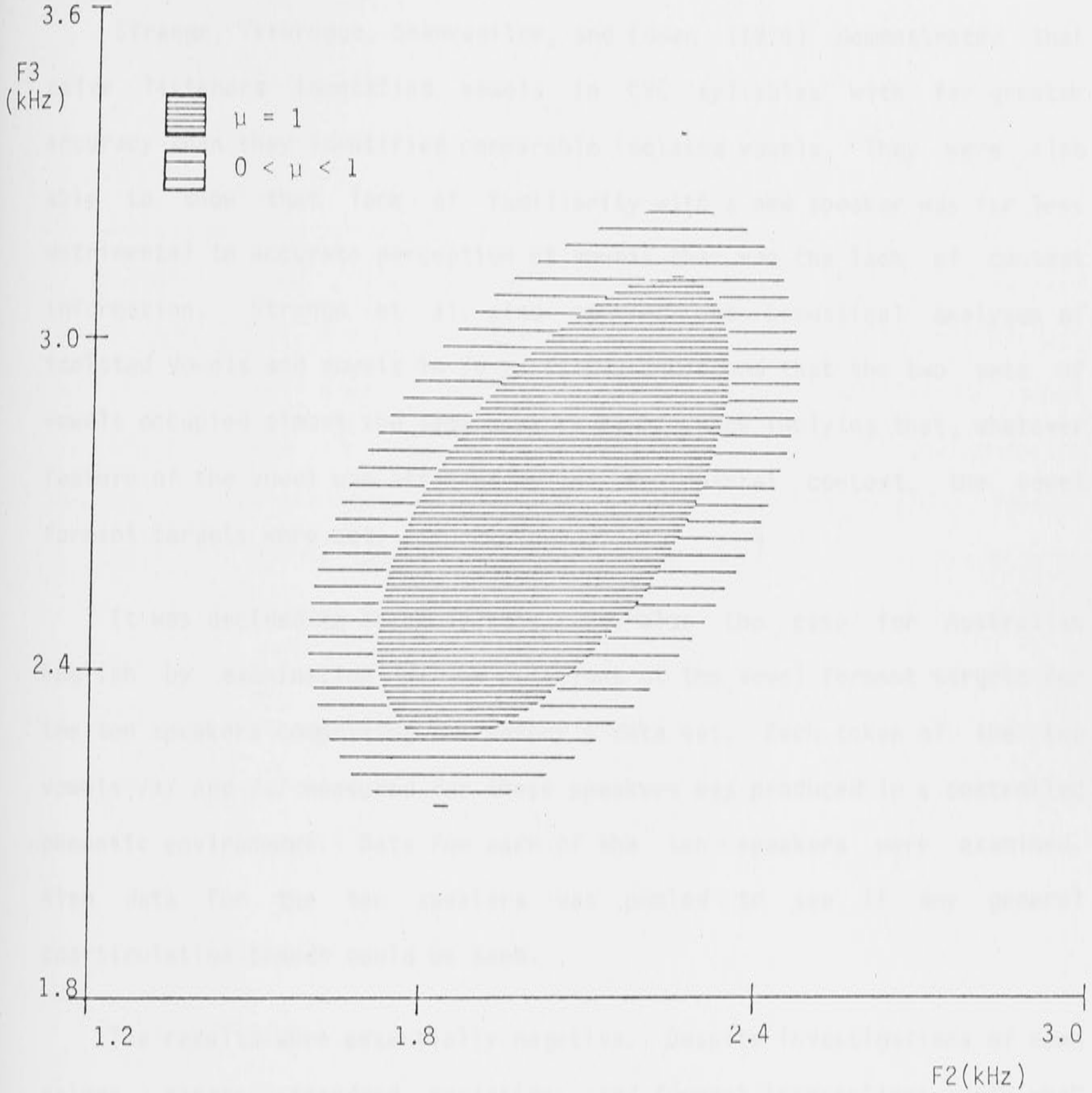


FIGURE 5.10: Two-dimensional fuzzy set giving the restrictions in F3-F2 space for the consistency condition for the vowel /i/ (male voices only).

5.6 EFFECT OF COARTICULATION ON THE POSITIONS OF VOWEL FORMANT TARGETS

Strange, Verbrugge, Shankweiler, and Edman (1976) demonstrated that naive listeners identified vowels in CVC syllables with far greater accuracy than they identified comparable isolated vowels. They were also able to show that lack of familiarity with a new speaker was far less detrimental to accurate perception of vowels than was the lack of context information. Strange et al. also carried out acoustical analyses of isolated vowels and vowels in /p-p/ frames and found that the two sets of vowels occupied almost the same area in F1-F2 space implying that, whatever feature of the vowel was affected by the consonantal context, the vowel formant targets were not.

It was decided to check if this was also the case for Australian English by examination of the positions of the vowel formant targets for the ten speakers comprising the author's data set. Each token of the two vowels /i/ and /ɔ/ measured for these speakers was produced in a controlled phonetic environment. Data for each of the ten speakers were examined. Also data for the ten speakers was pooled to see if any general coarticulation trends could be seen.

The results were essentially negative. Despite investigations of mean values, ranges, standard deviations, and formant interactions, only what were at best weak trends could be seen reflecting the influence of consonantal environment. For example, when examining data for individual speakers it was found that one of the strongest trends noticed was that Speaker 1 produced an /i/ which has a lower average second formant position when it occurred before a velar plosive /g/ or /k/ than when it occurred before any other plosive. For data pooled across all speakers the most significant effects were seen in the relative ranges of the vowels before a

variety of consonants. For example, the F2 ranges for the vowel /ɔ/ after the three types of plosives were as follows:

700-1250 Hz for /ɔ/ before /b/ or /p/

650-1250 Hz for /ɔ/ before /d/ or /t/

700-1175 Hz for /ɔ/ before /g/ or /k/

Examination of formant target positions for vowels in Oasa's data also failed to reveal any variation with target for the vowel targets of /æ/, /a/, /ɒ/, and /ɔ/. However the situation was very different for the vowel / / which showed very stable first formant but a widely varying second formant. For male speakers, /u/ before /l/ has an average second formant frequency of 918 Hz while for /u/ before a vowel in a final position or before /t/ the average male F2 value is 1532 Hz - a value which tallies with the Bernard average for /u/ in the /h-d/ frame. Thus although it must be concluded that vowel formant target positions do not generally display formant coarticulation effects there are certain special cases such as /u/ before /l/ where coarticulation effects become important.

5.7 A NOTE ON VOWELS IN CONTINUOUS SPEECH

Stevens (1968) showed that vowels in context are perceived more categorically than isolated vowels. Strange, Verbrugge, Shankweiler, and Edman (1976) showed that the existence of a consonant environment even when the consonant environment is not known in advance, significantly enhances the correct perception of vowels. Nevertheless Mermelstein (1978) has shown that the difference limens (DL) for the first two formant frequencies are significantly larger for the vowels in CVC contexts than those for steady-state vowels, with the DL for the second formant being larger in the direction of the expected shift due to consonantal coarticulation.

Presumably vowels in rapid conversational speech should display this effect even more than vowels in citation form CVC syllables. Such an effect no doubt accounts for the slightly different standard deviations associated with vowel formants in the author's data and the vowel formants in the Bernard and Oasa data sets and also the slightly different data ranges associated with the second formant of the vowel in certain contexts. (The effect is illustrated in Figure 5.1.)

Another discrepancy between Bernard's and the author's data which could be due to conversational speech being further removed from isolated vowel production than are citation form words is the fact that the average value (for male speakers) of F2 of /i/ is lower and the average value of F2 of /i/ is higher in the author's than in Bernard's data. This could be explained by Stevens and House's (1963) finding that the F2 of vowels in dynamic speech are in a more central position (i.e. a position closer to F2 of the neutral vowel /a/ with F2 at approximately 1400 Hz) than vowels produced in isolation.

The discussion in this section and in the previous section forces one to the conclusion that although vowels in context are more easily recognized, such an effect is certainly not due to enhanced separations of the formant positions for each vowel in formant space. This implies that it is necessary to investigate other vowel parameters to explain the enhanced perception of vowels in context. One parameter which could well yield coarticulatory results in vowel duration. This parameter will be considered in the next section. Other ways in which context helps define the vowel are coarticulation effects seen within consonants. This possibility will be explored within the next chapter.

5.8 VOWEL DURATION

Bernard (1967b) noted that in his data subjects often had the same vowel targets for the following pairs of vowels: /i/-/ɪ/, /a/-/ʌ/, and /ʊ/-/ɔ/ and yet these pairs can always be distinguished auditorily when they occur in the /h-d/ frame. He investigated this effect by running perceptual tests on artificially lengthened short vowels and artificially shortened long vowels, and thus found that the duration was the most important factor in distinguishing between these pairs of vowels in Australian English speech when the vowels are produced in stressed situations.

But it is common knowledge that the length of the vowel sound depends on where the vowel occurs in a word, whether it is stressed or not, and the speaking rate of the speaker (Klatt, 1976). Thus if duration is an important factor it cannot be the absolute duration of the vowel which is important but a duration measure that is coded in a complex manner which allows for all the effects mentioned.

Umeda (1975) conducted a study of the duration of American English vowels extracted from readings of long paragraphs by three speakers and discovered that the important factors affecting duration were position, the nature of the consonant following the vowel, and whether the vowel occurs in a content or a function word. Umeda proposed the following formula for vowel duration, T:

$$T = T_0 + S(K_1 + K_2 * C)$$

where T_0, K_1, and K_2 are specific to a particular vowel, S is a factor accounting for positional constraints, sentence stress, word prominence, and speech rate, and C is a factor to account for the consonantal condition.

The use of such a formula poses some problems for automatic recognition algorithms. An estimate of C requires that the consonant following the vowel be at least partially recognized. This supports the proposal that processing of different sounds should be done in parallel with rough recognition of one sound taking place while the fine recognition of the preceding sound is being completed. To estimate Umeda's constant C it is necessary to know to what class of sounds (voiceless plosive, nasal, voiced fricative, etc.) the consonant following the vowel belongs. An estimate of the factor S would be harder to obtain and would probably have to be approximated from measurements of speech rate (number of phonemes/unit time) and stress correlates such as pitch and amplitude. Once the two factors, S and C, are estimated and the duration, T, of the vowel is known it is still impossible to obtain values for T_0, K_1, or K_2 and hence decide what the vowel in question is. However, given that one knows from formant measurements that the unknown vowel can only be one of a limited set of vowels, one can try substituting the constants T_0, K_1, and K_2 for each one of the set of possibles into the formula to see for which vowel the right hand side of the equation is in closest agreement with the measured duration, T. On the other hand, given that the formula can at best be approximate and the estimate of S at best crude, the use of a look-up table is probably more appropriate.

5.9 AN ALGORITHM FOR VOWEL RECOGNITION

From consideration of the preceding sections an algorithm for the recognition of vowels in Australian English can be produced. The algorithm presented here is, like the algorithms presented in preceding chapters, a fuzzy algorithm based on the analysis of experimental data. The result of the algorithm thus gives a list of weighted possible identities for the unknown vowel.

Using Zadeh's (1976) composite function for a complex, imprecise concept:

$$Q \triangleq (U,B,A)$$

we are in this case considering a universe of all possible acoustic patterns which can be interpreted by a foreign phonetician as legitimate basic patterns of some language which is not his native language. B is the set of all vowels in the language being recognized (here we will restrict B to the set of monophthongs in Australian English speech; an adequate description of diphthongs would require formant transition information, and relative durational information about the components of the diphthong). A is a set of possible answers or evidence statements about the vowels.

The core of the fuzzy algorithm for vowel recognition consists of a set of statements or rules describing sufficient attributes of all the members of the set, B, of Australian English monophthongs. If an acoustic pattern which has already been classed as a possible vowel has, to a reasonable approximation, all the specified attributes of one of the members of the set, B, it will be classed as that member of the set and will be given a fuzzy recognition rating equal to the lowest degree of compatibility (a number between 0 and 1) between a component attribute of a pattern described in the algorithm core and the corresponding attribute of the pattern being recognized.

In previous sections of this chapter it was noted that successful perceptual categorization of a vowel involves allowance for variation of the positions of the formants with the sex of the speaker (or, more correctly variation with the length of the vocal tract of the speaker). Also as the system becomes more familiar with (i.e. is more regularly exposed to) the speaker's voice, vowel recognition improves. Knowledge of

the factors influencing vowel duration can change, in a human perceiver, vowel recognition from being continuous perception to categorical perception - this seems to be the case in Bernard's experiment anyway (1967b). These effects must be included in a recognition algorithm if that algorithm is to function in a manner compatible with human speech perception.

So far, normalization for the sex of the speaker has been included in the algorithm but on-going adaptation to a speaker has not. Duration conditions are only included for vowels in isolated words.

Let us consider then the core statements of the algorithm:

/i/ = (high front vowel)*(F2-F3 compatibility condition)
* ((i-duration condition) + (F2-F3 uniqueness condition for i))

/ɪ/ = (fairly high front vowel)*(F2-F3 compatibility condition)
*(ɪ-duration condition)

/ɛ/ = (medium front vowel)*(F2-F3 compatibility condition)
* ((ɛ-duration condition) + (F2-F3 uniqueness condition for ɛ))

/æ/ = (medium-low front vowel)*(F2-F3 compatibility condition)
* ((æ-duration condition) + (F2-F3 uniqueness condition for æ))

/a/ = (low front vowel)*(F3-F3 compatibility condition)
* (a-duration condition)

/ɜ/ = (medium central vowel)*(F2 in central frequency range)
*(ɜ-duration condition)

/ʌ/ = (low central vowel)*(low-medium position of F2)
*(ʌ-duration condition)

/ɒ/ = (low back vowel)*(F1-F2 compatibility condition)
*((ɒ-duration condition) + (F1-F2 uniqueness condition for ɒ))

/ɔ/ = (medium-low back vowel)*(F1-F2 compatibility condition)

* ((ɔ-duration condition) + (F1-F2 uniqueness condition for ɔ))

/ʊ/ = (medium-high back vowel)*(F1-F2 compatibility condition)

* ((ʊ-duration condition) + (F1-F2 uniqueness condition for ʊ))

/ɑ/ = (low F1)*(F2 dependent on following consonant)

*(ɑ-duration condition)

Let us consider the first statement above (i.e. the rule for /i/). This is a summary statement of the attributes which a potential vowel must have if it is to be classed as the vowel /i/. For actual testing of an unknown vowel we need a detailed numerical description of each of the component attributes. Such a description for a high front vowel would require a low first formant, and a high second formant which was dependent on the sex of the speaker.

The labels for the fuzzy restrictions relating to F1 for /i/ are:

$\{l_1, a_1, h_1\}$

and the corresponding vector of breakpoints measured in Hz is:

{150, 200, 350, 400}.

The labels for the fuzzy restrictions relating to F2 for /i/ are:

$\{l_2, a_2, b_2, c_2, h_2\}$

and the corresponding vector of break-points in Hz is:

{1400, 1800, 2350, 2500, 3000, 3150}

Then the rule for high front vowel is:

$HFV(x) = a_1(X * a_2 + \neg X * c_2)$

where X is a parameter which takes the value 1 if the speaker is male and the value 0 if the speaker is female, and HFV is written as a function of x to indicate that it depends on the sex of the speaker.

The F2-F3 consistency condition requires the point in F2-F3 space for the vowel being tested to have a high membership in the fuzzy set illustrated in Figure 5.10.

The F2-F3 uniqueness condition requires that the point in F3-F2 space of the vowel being tested lie in that region in which only points corresponding to /i/ are found. The fuzzy set used to test this condition is derived from Figure 5.8. This uniqueness condition is only useful in a limited number of cases.

The duration condition depends on many parameters as was discussed in Section 5.8. As was stated above only duration information for isolated vowels is sufficiently understood for incorporation into these rules. In other cases e.g. conversational speech, the duration condition is replaced by a number, usually 0.8.

For a vowel to be scored as a probable /i/ it must have fuzzy membership between 0.5 and 1 in all the components of the /i/ equation.

For the estimation of the likelihood of the unknown vowel being another vowel the procedure is essentially similar to the one described above for /i/.

Overall, the vowel recognition achieves 74% correct recognition. On conversational speech the recognition rate is somewhat lower at 68%, and on recognition of vowels in isolated words it is somewhat higher at 84%. It should be noted however, that 96% of the incorrect classifications are 'near misses'. Also the algorithm achieves considerably higher recognition scores for some speakers than for others, thus highlighting the need for on-going adaptation to each speaker.

5.10 CONCLUSION

In this chapter an algorithm for vowel recognition has been presented which is novel in that it does not depend on pre-training of the machine to each new speaker. Another novel feature of the algorithm is the use of consistency conditions between the formants. Such a condition helps to eliminate unlikely possibilities.

Chapter 6

COARTICULATION, JUNCTURE AND THE RECOGNITION OF PLOSIVE CONSONANTS IN AUSTRALIAN ENGLISH

6.1 THE PROBLEMS INVESTIGATED IN THIS CHAPTER

In Romance languages there is a noticeable aural effect whereby the sound at the end of one word influences the sound at the beginning of the following word (Malmberg, 1955). In English such an effect is not noticeable. Does this mean that there is no coarticulation across word boundaries in English? Apparently it does not as it has been stated that, at least for American English, this is one of the major problems for continuous speech recognition (Klatt, 1979; Klatt and Stevens, 1973). However there is little research on what actually does happen at word boundaries. Junctural situations associated with major grammatical boundaries are often characterized by long pauses and marked changes in pitch, but at the majority of word boundaries such prosodic effects do not occur. On casual inspection of acoustic waveforms of continuous speech, word boundary positions cannot be detected. Does this mean that the patterns of coarticulation the VCV sequence /itɔ/ are different if the sequence is extracted from the combination 'heat ought' than if it is extracted from the combination 'he taught'? Questions such as this are of great interest in the development of a consonant recognition algorithm. In Chapters 3 and 4 it was shown that consonant recognition could be enhanced by the use of coarticulatory information. But if 'normal' coarticulation is disrupted by the presence of a word boundary, schemes of consonant recognition which rely heavily on coarticulatory effects become of very

limited usefulness. Admittedly word boundaries do not seem to affect the results for recognition of the plosive consonants in Italian by the method described in Chapter 4 but, as was indicated above, coarticulation across word boundaries in Italian can definitely be heard. Also if it can be shown that coarticulation is unaffected by the presence of word boundaries, this will imply that it is unaffected by the presence of intraword syllable boundaries and this will vindicate the use of coarticulation-dependent consonant recognition algorithms in a wide variety of situations.

As well as providing a means of investigation of the interaction of coarticulatory and word boundary (hereinafter somewhat loosely termed 'juncture') phonomena, the experiment described in this chapter permits an investigation of which effects, if any, reflect the presence of the juncture at the phonetic (as opposed to the prosodic) level.

In this chapter an experiment is described in which pairs of VCV combinations, extracted from conversational speech are compared. In the first member of the pair a word boundary occurred after the consonant (e.g. /itɔ/ extracted from 'heat ought') and in the second the word boundary occurred after the first vowel (e.g. /itɔ/ extracted from 'he taught'). Data from ten speakers, five male and five female, were used. For each speaker forty-eight VCV combinations were examined. The consonants could be any of the six plosive consonants and the vowels were either /i/ or /ɔ/, the foremost and backmost vowels in Australian English respectively.

Summarizing, the main aims of this chapter may be stated as follows:

(a) To develop a recognition algorithm for the plosive consonants in Australian English.

(b) To discover to what extent such an algorithm can be made speaker independent.

(c) To examine the patterns of coarticulation in VCV utterances in Australian English.

(d) To see how the presence of a word boundary affects these coarticulation effects.

(e) To examine the phonetic correlates of juncture.

(f) To see whether or not all speakers display coarticulatory and juncture effects in the same way.

(g) To examine the possibility of features reflecting both juncture and coarticulation.

6.2 JUNCTURE AND ITS POSSIBLE INTERACTION WITH COARTICULATION

Studies on the interaction of juncture and coarticulation have been limited, the most comprehensive being that of Su et al. (1975) Hammarberg (1975) variations in the nasal consonant /m/ caused by the following vowel when a variety of juncture boundaries occurred between the two sounds. Not surprisingly they found that low-level junctures (i.e. where there is no change in grammatical structure) were marked by short pauses and did not disrupt nasal-vowel coarticulation. Higher-level junctures were found to be often marked by longer pauses, and concomitant reduction in nasal-vowel coarticulation was generally observed. However, it was also found that different speakers often had diagnostic means of signalling higher-level junctures. Thus some speakers would not pause at such junctural positions but would display reduced coarticulation, while others did the reverse.

These findings of Su et al. (1975) confirmed other findings on nasal-vowel coarticulation at junctural boundary by McClean (1973) and Moll and Daniloff (1971).

There have been several studies of junctural phenomena. Notable among these have been the work of Malmberg (1955), Lehiste (1960), and more recently Christie (1974, 1977). Malmberg constructed synthetic speech versions of the three nonsense VCV combinations, /ipi/, /odo/ and /aga/. In each version only one of the vowels had transitions towards the consonant attached to it. Listeners presented with these disyllables and asked to nominate to which vowel the consonant was attached always said that it was attached to the vowel with the transitions provided the period of silence between the vowels exceeded a certain minimum duration. If the silent period was less than this minimum duration the consonant was always said to be attached to the second vowel. As will be seen later, the silent interval in VCV combinations extracted from conversational speech is often about the same as Malmberg's minimum duration ($\sim$ 20-40 msec). In these cases one would expect the consonant to be heard as being attached to the following vowel regardless of the position of a word boundary.

Lehiste (1960), in her extensive study of juncture as it occurs in natural speech, had three male speakers read twenty-five pairs of phrases such as 'it sprays - it's praise' and also read sentences containing these phrases. Randomly ordered lists of the phrases were then presented to forty listeners who correctly identified them about 80% of the time. Spectrographic analysis of these contrastive word pairs revealed the following generalities:

(1) Word-initial, post-junctural allophones of almost all phonemes are considerably longer than either medial or pre-junctural allophones

(unless drawling occurs).

(2) An initial allophone is characterised by a rather rapid increase in acoustic energy compared to the energy decay for an allophone in a word-final position.

(3) American English voiceless stops are aspirated but a non-initial stop is usually unaspirated.

(4) Initial voiced stops are considerably longer than initial voiceless stops.

(5) Initial vowels may either start with a gradual rise in energy or with a glottal stop followed by an immediate onset of full energy.

(6) Final vowels are lengthened and decay gradually in energy.

(7) The formant position of vowels can play a part in determining whether a vowel is in an initial or medial position.

(8) There is a certain amount of variability in the junctural cues used by different speakers.

Christie (1974), claiming that it was impossible to deduce from Lehiste's work (1960) the relative importance of various cues in signalling juncture, used, like Malmberg (1955), synthetic speech to investigate various junctural cues in a systematic fashion. Investigating cues for a word boundary either before or after /s/ in the sequence /asta/ he found that the presence of aspiration in a voiceless stop is a strong cue for its being in a post-junctural position, but a very short silence interval before the stop combined with the presence of aspiration will yield ramdom responses. Continuing his investigations using synthetic speech Christie

(1975) concluded that there are several redundant cues for juncture and that while not all cues are necessary for the perception of juncture, one cue alone is often not sufficient for its perception.

It is evident, then, that there are several phonetic cues to the presence of juncture. Admittedly some of these cues are relative durational cues which are hard to measure automatically as one first has to know the general speaking rate of the speaker. Also it seems that unless there is at least a certain minimum silence between words (not necessarily long enough to be perceptible auditorily) the position of the word boundary are impossible to locate phonetically with any certainty.

In the experiment described in this chapter the junctural cues discussed above will be investigated, particularly with a view to seeing which cues are easy to detect automatically and finding with what degree of confidence a juncture can be postulated if such cues are present.

6.3 PLOSIVE IDENTIFICATION - WITH OR WITHOUT COARTICULATION?

In Chapter 4, the literature relating to coarticulation, particularly coarticulation between vowels and plosive consonants, was reviewed in some detail. Here only a few additional works will be considered, particularly as they relate to the specific questions investigated in this chapter and listed in Section 6.1.

The recent work of Stevens and Blumstein (Stevens and Blumstein, 1978; Blumstein and Stevens, 1979, 1980) has demonstrated that there is good reason to believe that plosive consonants are primarily perceived by an evaluation of the gross shape of spectrum of the consonantal release in CV situations or consonantal closure in VC situations. They nominate three

types of spectral shape - diffuse-falling, diffuse-rising, and compact - and proceed to show that in approximately 85% of the 900 CV syllables they examined these gross spectral shapes successfully described labial, alveolar, and velar plosive release spectra respectively. Furthermore, in a series of perception tests using synthetic speech Stevens and Blumstein showed that stimuli as short as 10-20 msec sampled from the onset of the consonant-vowel syllable can be reliably classified according to place of articulation. They also have shown that if gross onset spectral shape information is weak or ambiguous a plosive consonant can be identified from coarticulation-dependent properties such as transition movement. They postulate that coarticulation dependent measures provide a secondary means of plosive identification.

6.4 METHOD

The experiment designed by the author to investigate coarticulation, juncture, and plosive consonant phenomena is described in the following.

Lists of two-word sequences were prepared. Each of these two-word sequences was of either two forms:

(1) The first word ended in a VC combination and the second word began with a V; where the vowels could be either /i/ or /ɔ/ and the consonant was one of the six plosives e.g. 'heat ought', 'morgue awful';

(2) The first word ended in V and the second word began with a CV combination where the vowels could be either /i/ or /ɔ/ and the consonant was one of the six plosives e.g. 'he taught', 'more gory'.

Thus with the two vowels and six plosive consonants and two juncture

positions there were a total of forty-eight two-word combinations. The list of the two-word combinations is given in Table 6.1.

The five male and five female speakers all spoke standard educated Australian. They were from a geographically diverse area of Eastern Australia and ranged in age from twenty-four to fifty. Only one speaker (Speaker 3) was a non-native speaker of English. This speaker was originally from Germany, coming to Australia at age 12. However, in informal listening tests no-one judged him to be a non-native speaker of Australian English. It perhaps should be pointed out that Australian English is remarkably homogeneous geographically as regards pronunciation, the greatest variation occurring between people from different social and employment groups (Mitchell and Delbridge, 1965).

Each speaker was presented with the list of two-word sequences and instructed not to study the list but to immediately begin saying sentences containing the word sequences. It was impressed on the subjects that the sentences they produced were to be spoken at a conversational speed and that the semantic content of the sentences was of no particular importance. It was even suggested that slightly silly sentences would be perfectly acceptable. If a subject was having particular difficulty in producing a sentence containing any given two-word sequence he was told not to waste time over it but to say something such as 'I can't think of a sentence with --- --- in it'. All this was to keep the subject speaking at as conversational a rate as possible. This aim was largely achieved.

Experimental studies on conversational speech are difficult as it is hard to exercise any control over the many interacting parameters and still require the characteristic spontaneity of conversational speech. In the experiment described here phonetic and junctural contexts were controlled

TWO WORD COMBINATION	VCV SEQUENCE	TWO WORD COMBINATION	VCV SEQUENCE
glebe eating	/ib i/	he beat	/i bi/
glebe ought	/ib ɔ/	he bought	/i bɔ/
deed easily	/id i/	knee deep	/i di/
seed ought	/id ɔ/	sea daughter	/i dɔ/
fatigue easily	/ig i/	sea geese	/i gi/
fatigue ought	/ig ɔ/	he gored	/i gɔ/
sorb eating	/ɔb i/	raw beef	/ɔ bi/
absorb awful	/ɔb ɔ/	bore bought	/ɔ bɔ/
bored easily	/ɔd i/	law deed	/ɔ di/
board ought	/ɔd ɔ/	more doors	/ɔ dɔ/
morgue easily	/ɔg i/	more geese	/ɔ gi/
morgue awful	/ɔg ɔ/	more gory	/ɔ gɔ/
keep eels	/ip i/	he peels	/i pi/
sleep awfully	/ip ɔ/	tea poured	/i pɔ/
eat eels	/it i/	he teaches	/i ti/
beat all	/it ɔ/	he taught	/i tɔ/
bleak eating	/ik i/	he keeps	/i ki/
seek all	/ik ɔ/	he caught	/i kɔ/
warp easily	/ɔp i/	more peas	/ɔ pi/
warp all	/ɔp ɔ/	more pork	/ɔ pɔ/
bought eels	/ɔt i/	more tea	/ɔ ti/
bought all	/ɔt ɔ/	more talk	/ɔ tɔ/
walk easily	/ɔk i/	law keeps	/ɔ ki/
talk awfully	/ɔk ɔ/	law caught	/ɔ kɔ/

TABLE 6.1: Two-word combinations and the VCV sequence extracted from them.

and almost always the speech of the subjects sounded (in the author's subjective opinion) to be at conversational rate. Only when the author herself attempted the task was there any degree of stiltedness! It is perhaps indicative of the relative newness of continuous speech recognition research that paradigms such as the one used in the experiment outlined above generally have not been systematically developed as a means of controlled study of conversational speech phenomena.

The sentences containing the two-word sequences were recorded on a Nakamichi 550 cassette recorder, using a Bayer microphone The required VCV tokens (also listed in Table 6.1) were excised from the sentences using a waveform editing routine (Millar, 1978). These VCV combinations were then analysed using the Interactive Laboratory System waveform analysis package. For the male voices the speech was sampled at 10 kHz and for the female voices the sampling rate was 16 kHz. For each token a 12-coefficient for male voices and a 16-coefficient for female voices linear prediction analysis was done using the autocorrelation technique. Plots of the spectral peaks derived from this analysis as a function of time provided a means of measuring the formants. The waveform, and waveform amplitude were plotted simultaneously with this spectral peak versus time display, see Figure 6.1. From these combined plots various timing measures could be made. The burst was also located approximately using these plots and the cursor. It was located more accurately by study of a three-dimensional display of the spectra surrounding the estimated region of the burst. See Figures 6.2 and 6.3.

The three main measurements for plosive consonants are measurements concerning the formant transitions into and out of the burst, the spectrum of the burst, and the various timing measures. Here we shall consider all three such measurements to see how they reflect coarticulation and juncture effects.

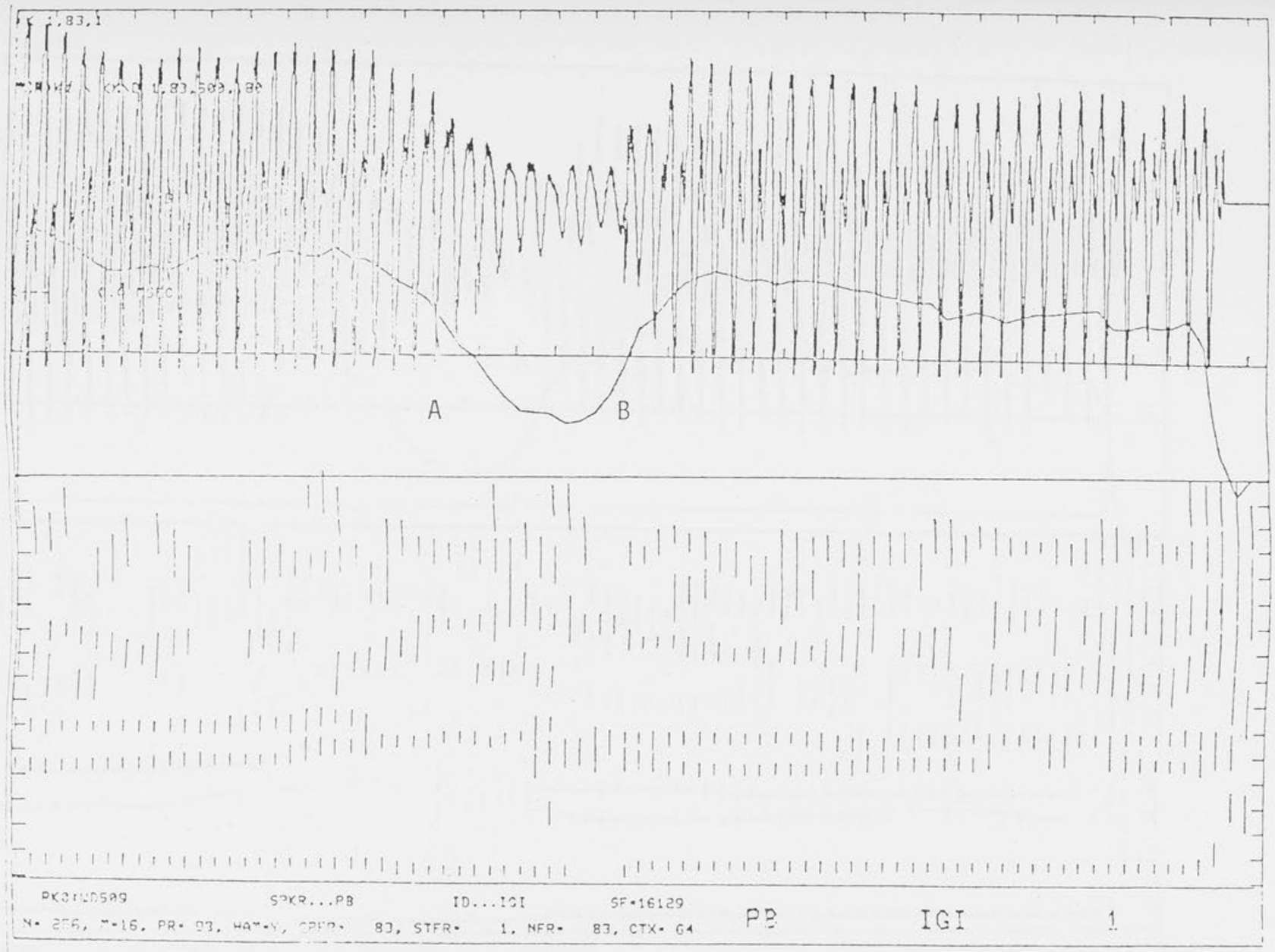


FIGURE 6.1(a): Time evolution of sampled data (top portion), signal energy (middle portion) and spectral peaks (lower portion) for the VCV sequence /ig i/ as produced by Speaker 6. (Display produced using ILS package.)

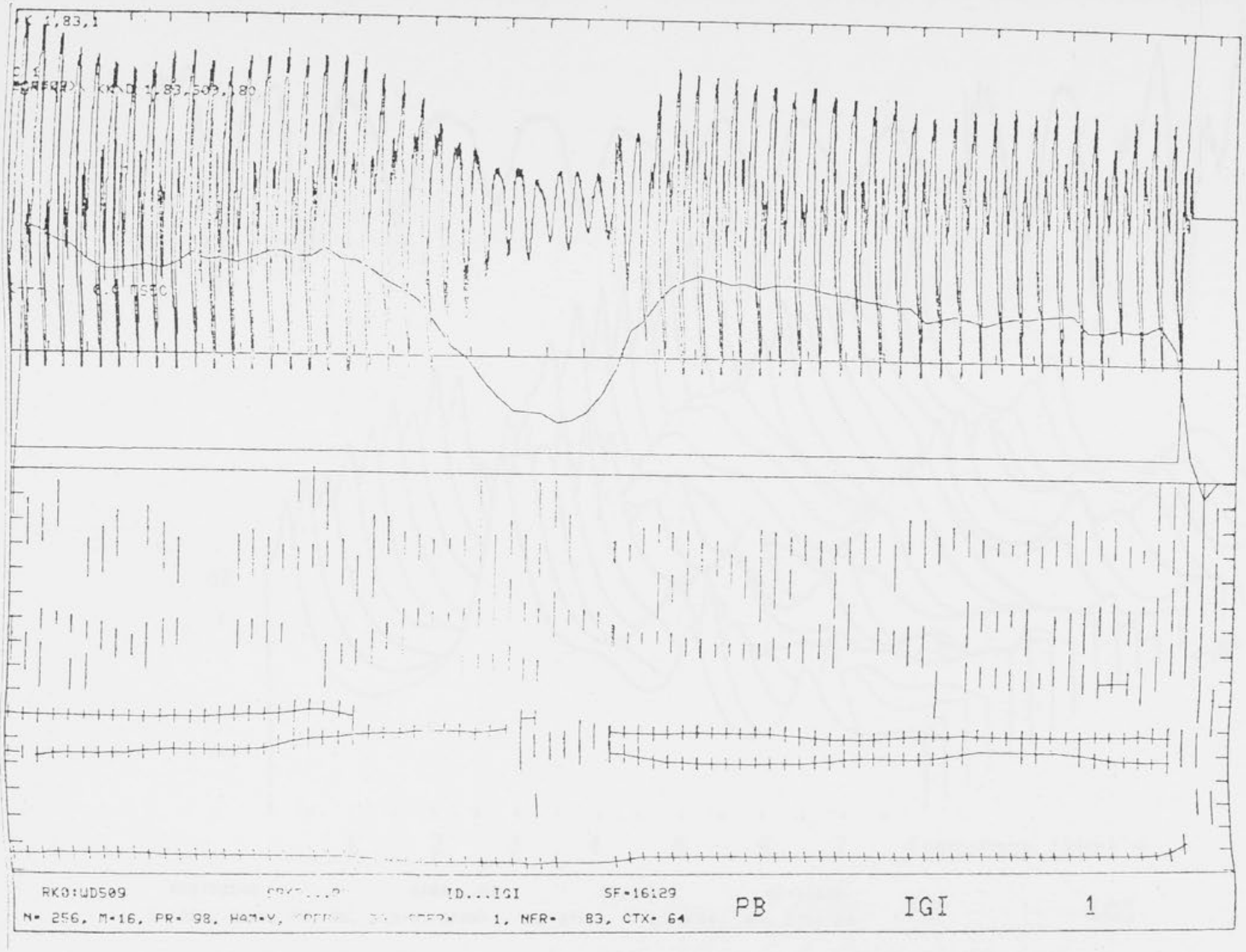


FIGURE 6.1(b): Same as Figure 6.1(a) except that formant tracks have been plotted by the automatic formant tracking routine. (Display produced using ILS package.)

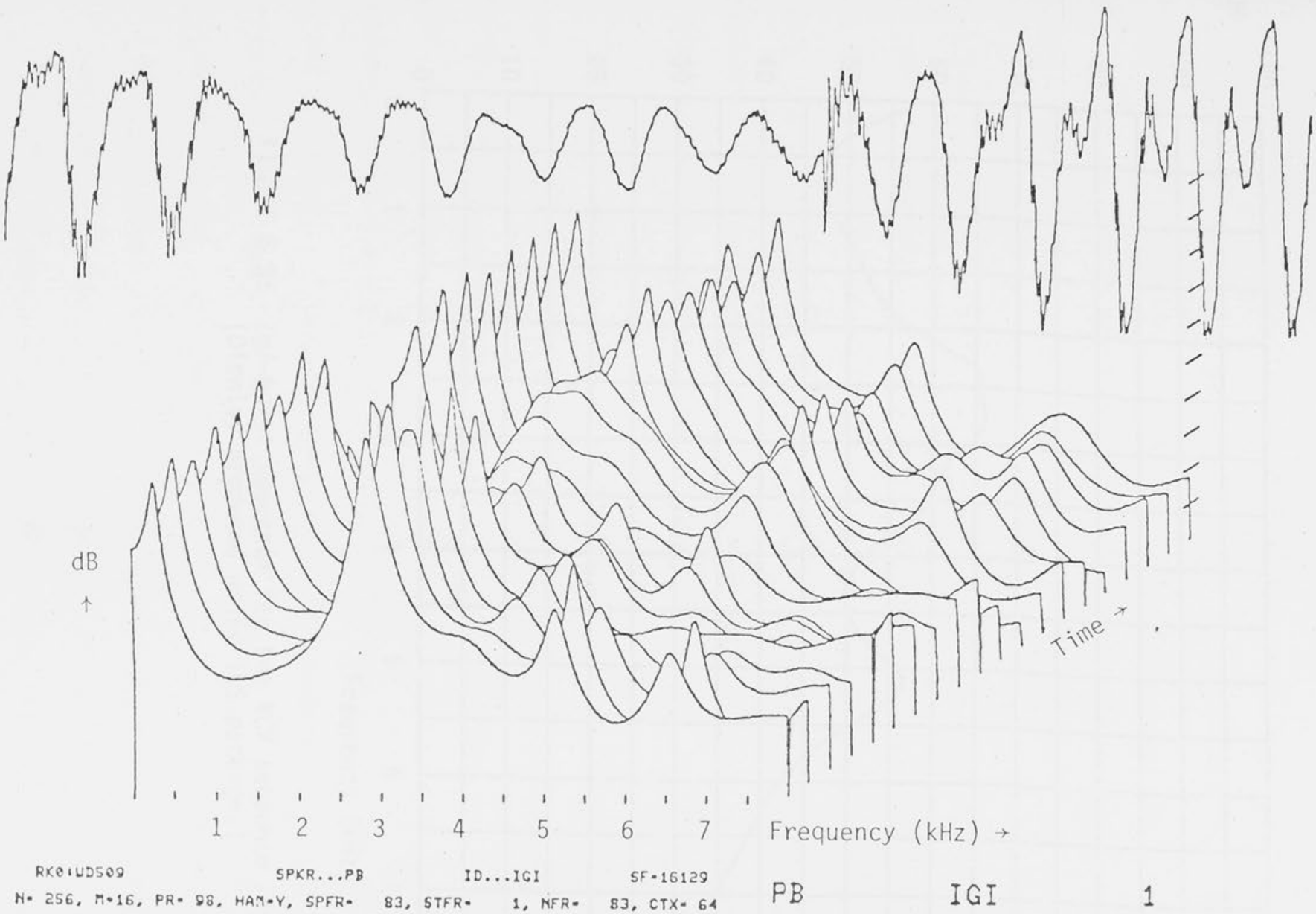


FIGURE 6.2: 3-D plot of spectra in the region of the /g/ burst for the VCV sequence /ig i/. Spectra calculated every 4 msec.
(Display produced using ILS package.)

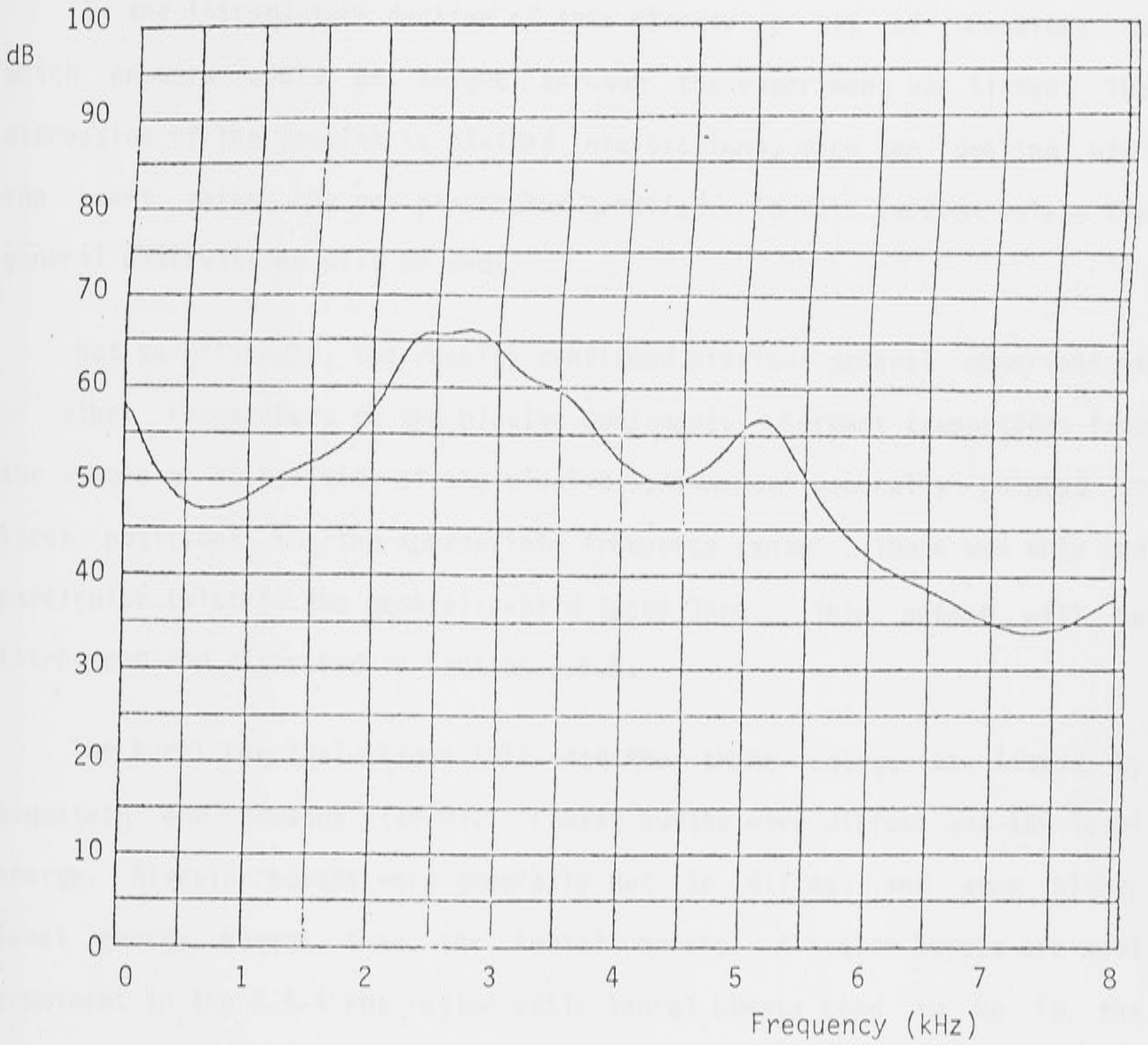


FIGURE 6.3: /g/-burst spectrum for the VCV sequence /ig i/.
(Display produced using ILS package.)

6.5 RESULTS

In the introductory section of this chapter a set of questions to which answers would be sought through the experiment was listed. The discussion of the results is divided into sections, each one dealing with the point raised by one particular question. In this section only a few general observations will be made.

Not surprisingly, the results confirmed previous general observations of other researchers on the plosive consonants. Formant transitions from the vowels on either side of the plosive consonants generally pointed to locus positions in the appropriate frequency ranges. There was only one particular twist to the generally-held locus lore. This effect will be introduced and discussed in Section 6.9.2.

The burst spectral shapes fell into the three categories listed by Blumstein and Stevens (1979). Labial bursts were diffuse and low level energy. Alveolar bursts were generally not so diffuse and were higher level energy bursts than the labial bursts. Alveolar bursts are most prominent in the 2.5-4 kHz region while labial bursts tend to be in the 1-3 kHz region. Velar bursts characteristically display two narrow-bandwidth peaks. The more prominent of these occurs in the 0.7-2.8 kHz region, and the (generally) smaller one occurs in the range 3.5-5 kHz. The exact position of these peaks is dependent on the nature of the surrounding vowels, an effect which will be described in detail in Section 6.9.3. It should be noted that the description of velar bursts as two-peak bursts is different from Blumstein and Stevens' (1979) description of these bursts. Typical examples of the three burst shapes are given in Figure 6.4 (a)-(c).

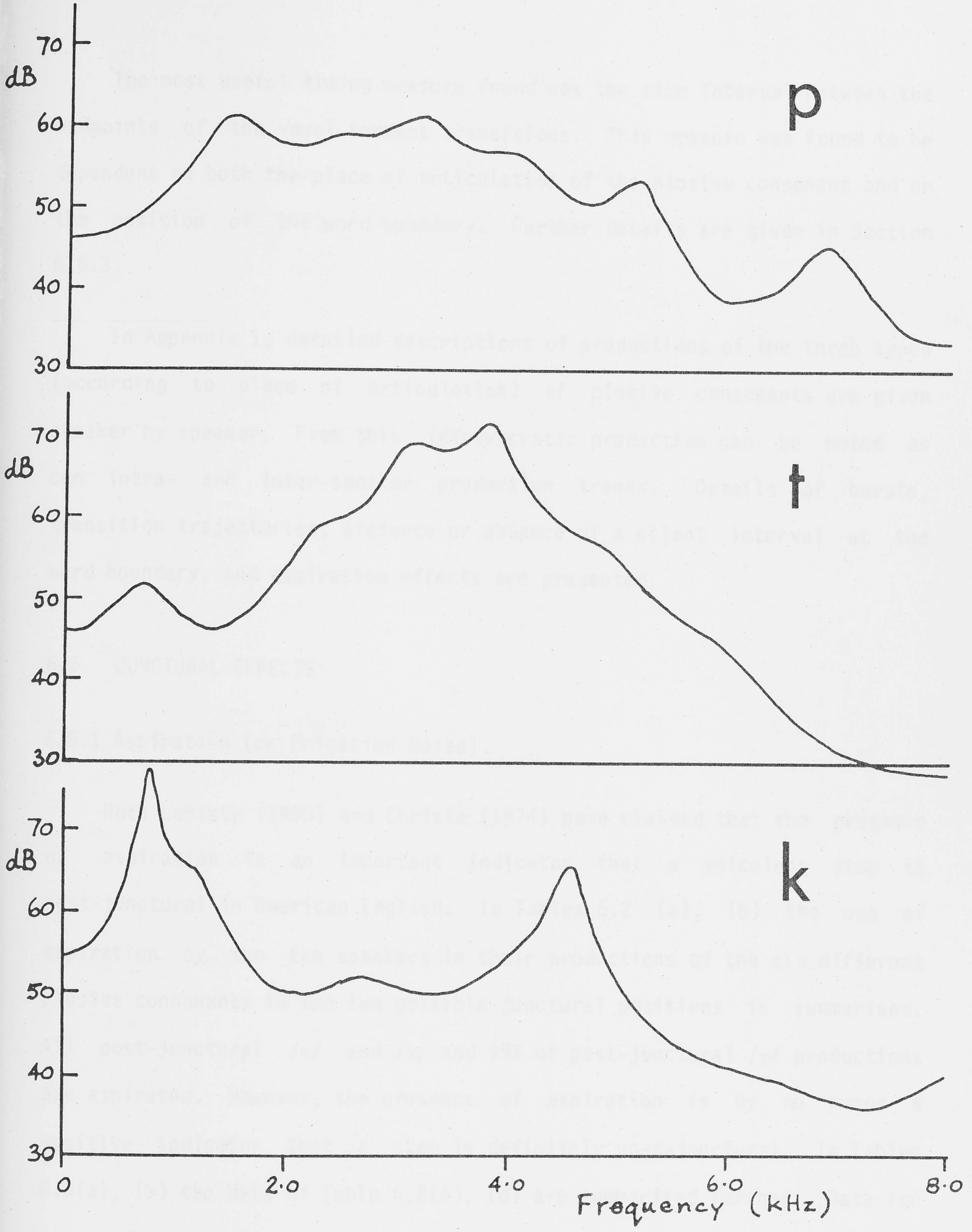


FIGURE 6.4: Typical examples if bilabial (top), alveolar (middle),and velar (bottom) bursts. These bursts were extracted from /ɔ-ɔ/ contexts and were produced by Speaker 6.

The most useful timing measure found was the time interval between the endpoints of the vowel formant transitions. This measure was found to be dependent on both the place of articulation of the plosive consonant and on the position of the word boundary. Further details are given in Section 6.6.3.

In Appendix 1, detailed descriptions of productions of the three types (according to place of articulation) of plosive consonants are given speaker by speaker. From this, idiosyncratic production can be noted as can intra- and inter-speaker production trends. Details of bursts, transition trajectories, presence or absence of a silent interval at the word boundary, and aspiration effects are presented.

6.6 JUNCTURAL EFFECTS

6.6.1 Aspiration (or Frication Noise)

Both Lehiste (1960) and Christe (1974) have claimed that the presence of aspiration is an important indicator that a voiceless stop is post-junctural in American English. In Tables 6.2 (a), (b) the use of aspiration by the ten speakers in their productions of the six different plosive consonants in the two possible junctural positions is summarised. All post-junctural /t/ and /k/ and 95% of post-junctural /p/ productions are aspirated. However, the presence of aspiration is by no means a positive indicator that a stop is definitely post-junctural. In Tables 6.3(a), (b) the data of Table 6.2(a), (b) are summarised further. Data for cases where aspiration is quite marked are presented as well as for cases where aspiration is at least seen by visual inspection of the waveform because the former represent the cases for which aspiration is aurally very recognizable.

	p		t		k		b		d		g	
SPEAKER	VC	V	V	CV	VC	V	V	CV	VC	V	V	CV	VC	V	V	CV
1	2		4		3		4		2		4		0		1		0		0		0		3	
2	2		4		2		4		4		4		0		1		0		2		1		3	
3	1		4		3		4		4		4		1		2		3		4		0		4	
4	4		4		4		4		3		4		1		4		0		4		1		3	
5	4		4		4		4		4		4		0		1		1		2		2		2	
6	4		4		4		4		4		4		0		2		1		1		1		3	
7	2		3		4		4		3		4		3		1		1		4		2		4	
8	3		4		4		4		3		4		0		4		0		4		3		4	
9	3		4		3		4		3		4		1		3		2		4		2		4	
10	2		3		3		4		3		4		0		0		1		2		1		3	

TABLE 6.2(a): Number of cases (out of a possible 4) for which aspiration is present in the production of the consonant. Data classed according to manner, place, and juncture position.

	p		t		k		b		d		g	
SPEAKER	VC V	V CV	VC V	V CV	VC V	V CV	VC V	V CV	VC V	V CV	VC V	V CV
1	2	4	1	4	2	4	0	0	0	0	0	2
2	1	4	2	4	4	4	0	0	0	0	0	2
3	1	4	2	4	4	4	1	1	2	4	0	4
4	3	4	4	4	2	4	0	1	0	4	1	3
5	1	4	3	4	4	4	0	1	1	1	0	1
6	2	4	4	4	3	4	0	0	0	0	0	3
7	1	3	4	4	3	4	1	0	1	4	2	4
8	2	4	4	4	3	4	0	3	0	4	2	4
9	1	4	3	4	3	4	0	1	2	4	1	4
10	0	0	1	4	2	4	0	0	0	1	0	1

TABLE 6.2(b): Number of cases (out of a possible 4) for which aspiration is very marked in the production of the consonant. Data classed according to manner, place, and juncture position.

	VC V	V CV
p	68	95
t	88	100
k	90	100
b	15	48
d	23	68
g	33	83

TABLE 6.3(a): Percentage of cases for which aspiration is present in the production of the consonant. Data averaged over the ten speakers and classed according to manner, place, and junctural position.

	VC V	V CV
p	35	88
t	70	100
k	75	100
b	5	18
d	15	55
g	15	70

TABLE 6.3(b): Percentage of cases for which aspiration is very marked in the production of the consonant. Data averaged over the ten speakers and classed according to manner, place, and junctural position.

From the Tables 6.2 and 6.3 it can be seen that the presence of aspiration depends on the manner, place of articulation, and junctural position. Voiceless stops are more heavily aspirated than voiced stops for any particular place of articulation and junctural position; velar stops are more heavily aspirated than alveolar stops which are in turn more heavily aspirated than labial stops given that voicing and junctural conditions are held constant; and post-junctural stops are more heavily aspirated than pre-junctural stops with the same manner and place of articulation. Perhaps the most surprising result is the high percentage of aspiration associated with the voiced stops, particularly the velar stops. It is clear that the presence or absence of aspiration can only be used as a guide to a juncture boundary if something is known about the place of articulation and the manner of voicing of the stop consonant and even then only as a very limited guide. The differences between the findings discussed here and those of Lehiste (1960) is probably mainly due to the fact that while Lehiste's subjects were reading the subjects in this experiment were speaking at a conversational rate.

Before leaving the subject of aspiration it should be noted that different speakers use aspiration in different ways in the production of various sounds. Some speakers, notably speakers 1, 2, 6 and 10, generally produce unaspirated /b/ and /d/ while other speakers notably speakers 4 and 8, produce unaspirated /b/ and /d/ in prejunctural positions, and aspirated /b/ and /d/ in postjunctural positions. And so on.

6.6.2 Silence

Another parameter which was investigated as a possible cue to the presence of a word boundary was the existence of a period of silence (i.e. the waveform amplitude fell below a desirable level) at the word boundary. As for the aspiration cue, we find that the silence cue is dependent on the manner of voicing and the place of articulation of the stop as well as on the juncture position. The results for the presence of silence are presented in Tables 6.4 and 6.5. Of course, the presence of silence will also depend to some extent on the semantic content of the sentence. For example, if the word following the word boundary is heavily emphasised there will often be a long period of silence at the word boundary regardless of whether the emphasised word begins with a consonant or not, and what the manner of voicing and place of coarticulation of that consonant is. It should also be noticed by reference to Table 6.4 that some speakers use silence in highly idiosyncratic ways. Speaker 7's production of labial stops are always accompanied by a period of silence, regardless of voicing manner, or juncture position. Other speakers, notably Speaker 1, 2, 4 and 9 rarely have any silence at word boundaries if the consonant is voiced.

6.6.3 Timing

It is to be expected, from a survey of previous studies on juncture and also intuitively that a timing parameter is most likely to give good distinctions between the two juncture cases. A parameter which shows this distinction quite well and which is reasonably easy to measure automatically is the time interval between the endpoints of vowel 1-consonant and consonant-vowel 2 formant transitions. More precisely this time interval, T, was defined as beginning at the point A

	p		t		k		b		d		g	
SPEAKER	VC V	V CV	VC V	V CV	VC V	V CV	VC V	V CV	VC V	V CV	VC V	V CV
1	3	4	2	4	4	4	0	3	0	0	0	1
2	0	3	1	4	2	4	0	0	0	0	1	0
3	2	4	1	4	2	3	1	2	1	2	0	4
4	2	1	2	2	2	4	1	0	0	0	1	1
5	4	4	0	3	1	2	2	2	0	1	0	1
6	2	3	0	4	4	4	0	3	0	1	0	2
7	4	4	1	4	4	4	4	4	0	4	2	3
8	3	4	3	4	4	4	1	4	0	3	1	3
9	4	4	2	4	3	3	1	2	0	0	1	1
10	4	3	4	4	4	4	3	3	2	0	2	2

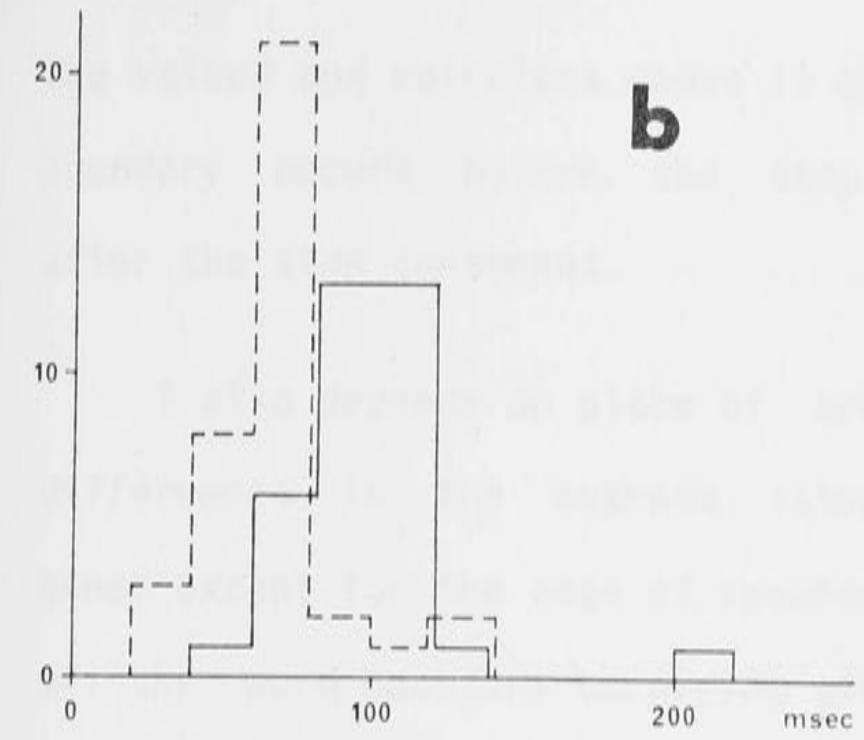
TABLE 6.4: Number of cases (out of a possible 4) for which a silence period exists (i.e. the waveform amplitude drops to zero) at a word boundary. Data classed according to manner, place, and juncture position.

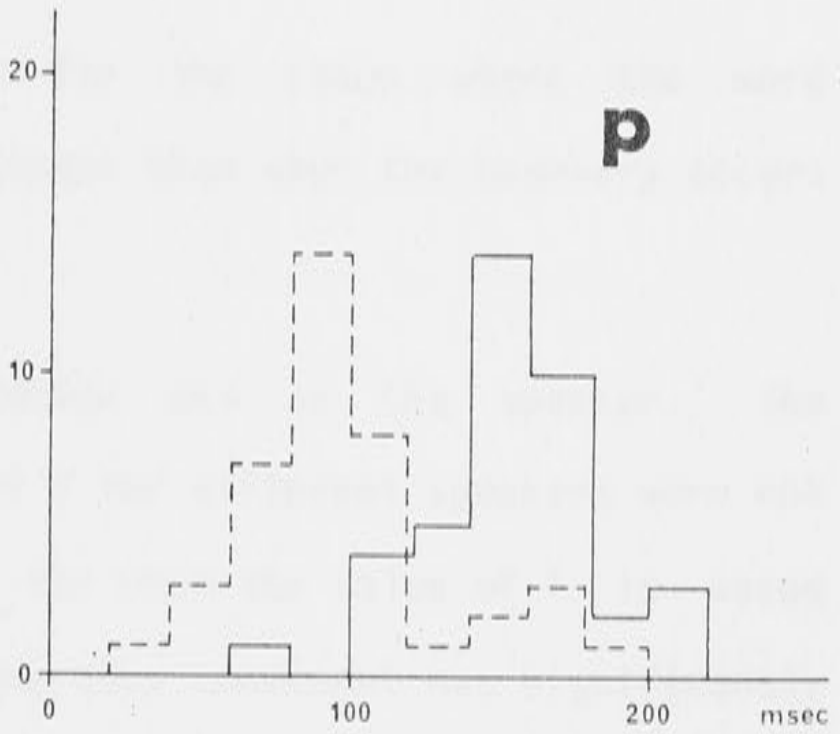
	VC V	V CV
p	70	85
t	40	93
k	75	90
b	33	58
d	8	28
g	20	45

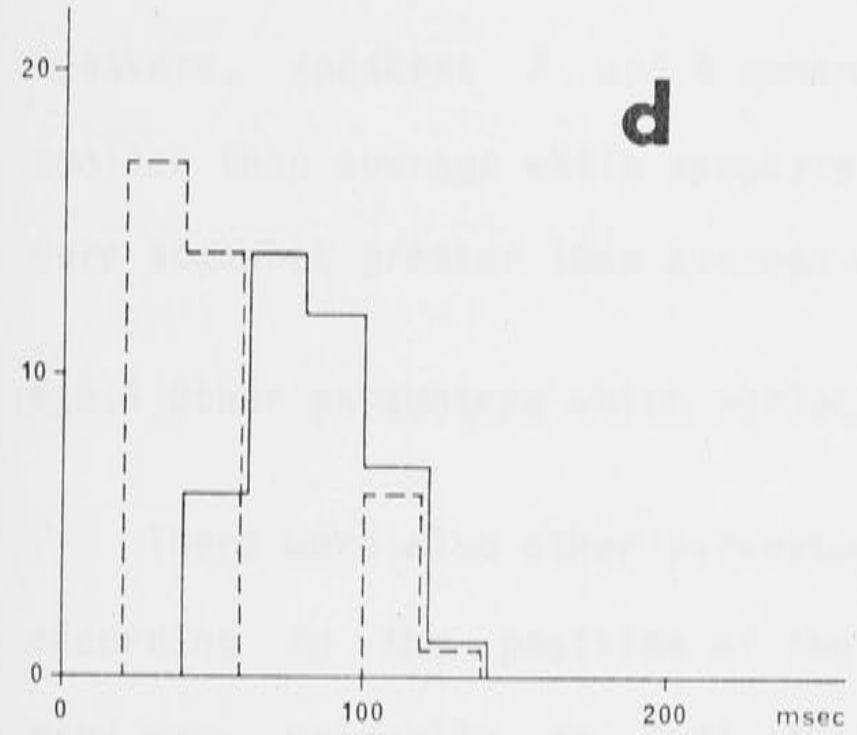
TABLE 6.5: Percentage of cases for which a silence period exists at a word boundary. Data averaged over the ten speakers, and classed according to manner, place, and junctural position.

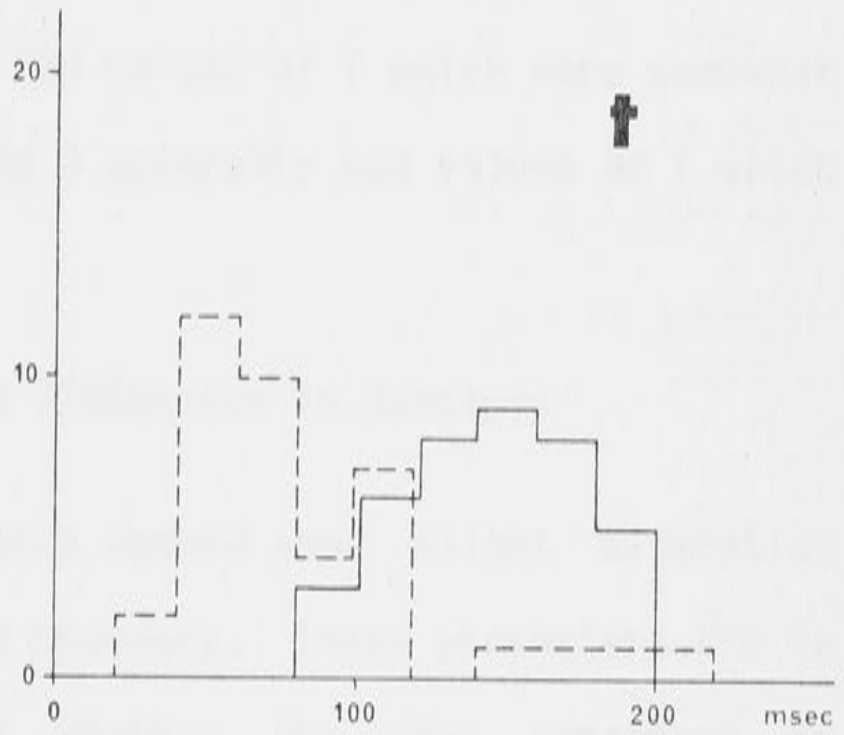
(as marked in Figure 6.1), the time corresponding to the minimum gradient as (measured from conventional zero) point of the waveform rms energy curve (middle display in Figure 6.1) in the region in which the rms energy decreases from its value for the steady state of the vowel to the closure for the stop consonant, and ending at the point, B (as marked on Figure 6.1), the time corresponding to the first point of maximum slope of the waveform energy curve in the region in which the rms energy curve increases after the plosive burst to its (much higher) value during the steady state of the vowel following the consonant. It should be noted that, regardless of juncture position, the decline in the waveform rms energy from the steady state of the vowel preceding the stop consonant to the closure before the consonantal release is always gradual while the increase in rms energy from just after the stop burst to its value for the steady state of the vowel following the consonant is rapid and the slope of the energy curve stays constant during most of this increase. It is for this reason that points A and B are not defined symmetrically.

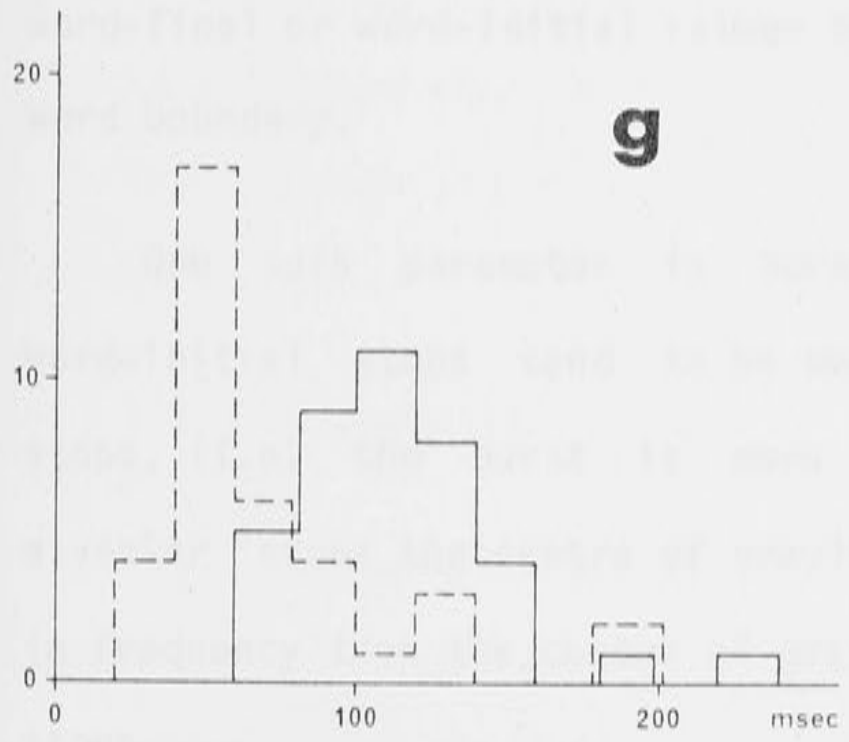
In Figure 6.5 histograms showing the number of times the timing parameter, T, falls in different ranges for different classes, are displayed. For any given stop consonant, T was, in most cases, larger for the cases where the word boundary occurred before the stop consonant than for cases where the word boundary occurred after the stop consonant. From Figure 6.5 it can also be seen that T is also dependent on whether or not the stop consonant is voiced; T being larger for voiceless stops than for voiced stops. This is not a surprising result as the parameter, T, incorporates VOT (voice onset time) as well as some measure of the word boundary timing. And it has been shown that VOT is larger for voiceless stops than for voiced stops (Lisker and Abramson, 1967). It is also interesting to note that the difference in the average value of T between











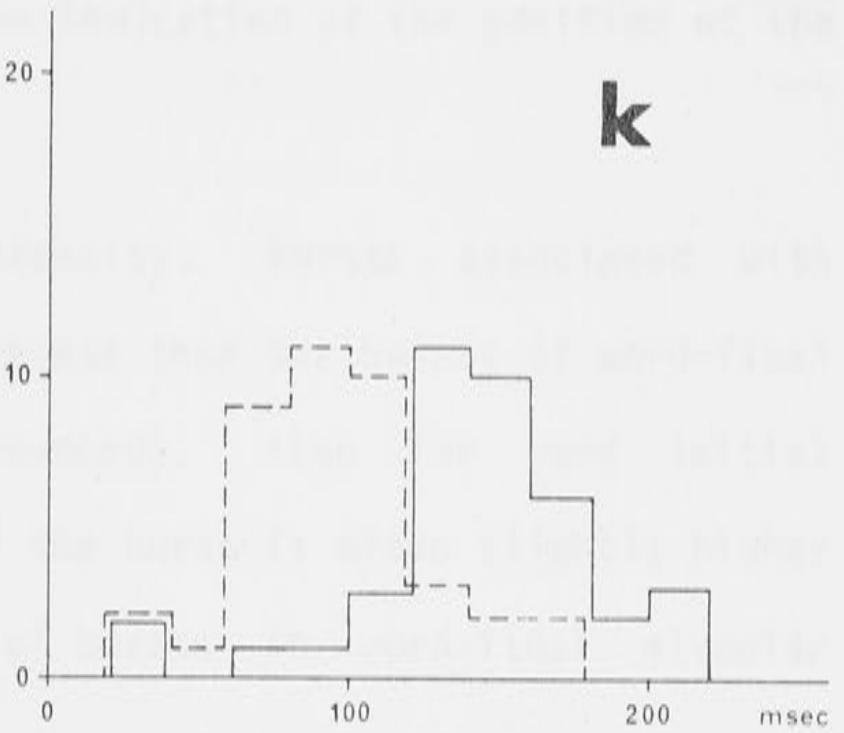


FIGURE 6.5: Histograms showing the distribution of the parameter T. Data associated with each plosive consonant shown separately. The dotted lines represent cases in which the plosive occurred in a /VC V/ situation and the full lines represent the cases in which it occurred in a /V CV/ situation.

the voiced and voiceless cases is greater for the cases where the word boundary occurs before the stop consonant than when the boundary occurs after the stop consonant.

T also depends on place of articulation and on the speaker. The differences in the average values of T for different speakers were not great except for the case of speaker 10, for whom the value of T in cases of the word boundary occurring after the stop consonant was significantly larger than the across-speaker average value of T. Among the other nine speakers, speakers 2 and 4 generally had values of T which were somewhat smaller than average while speakers 8 and 9 generally had values of T which were somewhat greater than average.

6.6.4 Other parameters which reflect the difference in juncture

There were also other parameters which showed some slight alteration according to the position of the word boundary. These parameters are in many ways primarily an indication of whether the stop consonant is word-final or word-initial rather than an indication of the position of the word boundary.

One such parameter is burst intensity. Bursts associated with word-initial stops tend to be more intense than the bursts of word-final stops, (i.e. the burst is more pronounced). Also for word initial alveolar stops the centre of gravity of the burst is often slightly higher in frequency than the centre of gravity of bursts in word-final alveolar stops.

Other effects include the fact that in 8% of all word-final stops there is no discernible burst whereas for word-initial stops less than 1% of all cases have no burst. Also, for word-final stops the burst spectra

are often (31% of all cases for which there is a burst) a superposition of the preceding vowel formant structure and the 'expected' burst spectrum. For word-initial stops this effect occurs in 5% of the cases. For all other word-initial stops a 'pure' burst spectrum was seen.

6.7 THE LISTENING EXPERIMENT

In the preceding section it was seen that there were several parameters which seemed to reflect to varying degrees the position of the word boundary in the VCV sequences considered. In an attempt to see if these factors were perceptually important in allowing a listener to judge the position of a word boundary, a group of listeners was presented with randomized lists of the VCV sequences produced by two speakers and asked to mark on an answer sheet where the word boundary had occurred.

6.7.1 Subjects

Seven, phonetically naive, Computer Science students voluntarily participated in the experiment. None of these participants had taken part in the recording experiment; nor were any of them familiar with the voices of the speakers who made the recordings. All participants claimed to have normal hearing.

6.7.2 The experimental set-up

The VCV sequences used in the listening experiment were extracted from conversational speech as described in Section 6.4. All started at the beginning of the steady state position of the first vowel and finished at the end of the steady state position of the second vowel. It was possible in some cases to tell what the sound preceding the first vowel or the sound following the second vowel was.

Answer sheets were prepared in which the two possibilities for each VCV sequence were listed, e.g. if the VCV sequence was /ɔg ɔ/, as extracted from the two-word combination 'morgue awful', the two possibilities were written as

/ɔg ɔ/ or /ɔ gɔ/

Subjects were asked to mark which version they heard. Thus subjects only had to concentrate on the position of the word boundary and not on identity of the sounds they were hearing. Subjects were instructed to mark an answer for every VCV sequence they heard. When they were uncertain of the position of the word boundary they were told to guess. Two subjects did not comply with this request so only the results for the five subjects who did mark every answer are considered below.

Twelve, randomly picked VCV sequence examples as produced by speaker 7 were used as a trial set to accustom subjects to the experiment. The experiment proper consisted of the presentations of the 48 (randomized) VCV sequences produced by speaker 5 (male), followed by the 48 (randomized) VCV sequences produced by speaker 9 (female). All the VCV sequences for one speaker were presented in one grouping as it was interesting to see if subjects did better as they became increasingly familiar with a speaker's voice. There was a short pause (about 4 seconds) between each VCV sequence and subjects were told when they would hear a new speaker's voice.

The recorded VCV sequences were played on a Revox taperecorder to subjects in a reasonably small classroom.

6.7.3 Results

The number of correct responses, out of a possible 48, of each listener for the two speakers are given in Table 6.6. In all but one case all subjects achieved scores of better than 50% correct responses. Although the number of correct responses is not very great it is clear that overall, the listeners are doing better than chance, because the probability that the total number of correct responses was due to chance is negligible being $O(10^{-4})$ according to normal distribution tables. In this case the normal distribution can be used to to approximate the result of this binomial experiment.

There are several other significant results for this experiment. For the remainder of this section where the word 'significant' is used, it is to be understood that the appropriate null hypothesis can be rejected using a t-test at a significance level of 0.1 or less. Where the symbols (*) occur it is to indicate that the null hypothesis involved can be rejected at a significance level of 0.025.

Some listeners tended to be more proficient at the task than others. Thus listener 3, who was the best performer overall, performed significantly better than listener 2(*), the worst performer, and listener 4. Listeners 1 and 5, the second-best performers, also performed significantly better than listener 2.

While the responses to each of the two speakers' voices, averaged across all listeners, were not significantly different, listener 1 achieved significantly (*) better results on speaker 5's speech than on speaker 9's speech. This raises the possibility that different speakers use a variety of (but not necessarily all possible) cues to signal juncture at the

		Listener				
		1	2	3	4	5
Speaker	5	34	23	31	28	28
	9	25	26	33	27	31

TABLE 6.6: Number of correct responses (out of a possible 48) for each of the five listeners to each of two speakers in the aural juncture location experiment.

phonetic level and that some listeners are responsive to some cues but not to others.

In informal discussion after the experiment all listeners claimed that they became increasingly adept at the task as they heard more examples from a particular speaker. Considering listeners' performance over the last third of the data sets for each speaker as opposed to their performance over the first two-thirds of the data sets, it is clear that listeners significantly adapt to a speaker.

Another point that the five listeners agreed on in informal discussion is that there were 'many' cases in which they had no doubt whatever about the position of the word boundary. Nevertheless, for speaker 5 there were only four cases in which all listeners correctly judged the position of the word boundary; and only one such case for speaker 9. For both speakers 5 and 9 there was one case each in which all listeners misjudged the position of the word boundary.

This certainty of listeners as to their correctness in many cases leads us to what were by far the most interesting results of this experiment. Although for each speaker there were 24 instances in which the word boundary occurred after the consonant and 24 instances in which the word boundary occurred before the consonant, the listeners responses favoured the word boundary occurring before the consonant in 75% of the responses to speaker 5's voice, and 70% of the responses to speaker 9's voice. Thus of the two types of mistakes it was possible to make, all listeners made the mistake of judging that the word boundary occurred before the consonant when it had actually occurred after the consonant, significantly (*) more often than they made the mistake of judging that the word boundary occurred after the consonant when it had in fact occurred

before the consonant.

This is a very interesting result in the light of Malmberg's (1955) finding that if the silent interval before the stop in his synthetic VCV syllables was very short, the stop was always heard as being attached to the second vowel. It seems that in rapid conversational-style speech where there is often no completely silent period, and where the time interval between the endpoints of vowel-consonant transitions is often as short as 20 msec, there is a strong possibility that listeners will hear a consonant as being attached to the vowel following it regardless of the position of the word boundary.

In summary, this experiment indicates that it is very difficult perceptually to reliably judge juncture boundaries from acoustic-phonetic information alone. Some listeners perform better than others at the task and all listeners improve their performance slightly as they adapt to the speaker's voice. In many cases a listener will feel confident that he has judged the junctural position correctly but this can be due to the tendency for consonants to sound as though they are attached to the following vowel when the period of silence before the release of the stop is short.

6.8 LOCATING JUNCTURE BOUNDARIES AUTOMATICALLY

From the preceding discussion it will be clear that the location of juncture boundaries from acoustic-phonetic information alone is not easy and cannot be done at all in many cases. Nevertheless, there are some cases in which a word boundary can be hypothesized with some degree of certainty. Before giving the fuzzy rules for these cases we should note that in the experiment described in Section 6.4 only trans-junctural VCV sequences were considered but that intraword VCV sequences were not.*

Nevertheless, from informal observations it is expected that the consonant in intraword VCV sequences will be more like a word-final consonant than a word-initial consonant. This implies that some parameters which seem from the results of Section 6.5 to indicate junctural position, only indicate it *if* a word boundary is present it occurs in some particular place. Thus, we have the following fuzzy rules which allow some statement about juncture positions:

For vowel-stop consonant-vowel sequences:

(1) If the consonant is voiceless and it is unaspirated then there is either no juncture within the VCV sequence or if juncture is present it occurs after the consonant with fuzzy membership 0.9.

(2) If the consonant is voiceless and there is no silence period, then there is either no juncture within the VCV sequence or if juncture is present it occurs after the consonant with fuzzy membership 0.8.

(3) If the consonant is voiced and aspirated then a juncture is hypothesized as occurring before the consonant with fuzzy membership 0.9.

(4) If the parameter T (defined in Section 6.6.3) appropriate to the prior classification of the consonant falls within the range of the appropriate fuzzy set, defined from the dotted line histograms of Figure 6.5, then a juncture is hypothesized as occurring before the

* No intraword VCV sequences were considered because it was impossible to find a complete (or even nearly-complete) set of words containing VCV sequences made up of the vowels /i/ and /ɔ/ and six plosive consonants without resorting to nonsense words and proper names. Nonsense words and proper names were avoided as it is likely that their production is accompanied by artificial stress and timing effects which do not reflect the general character of conversational speech.)

consonant with fuzzy membership equal to 0.9.

It is interesting to note that rule number 4 allows the assignment of the word boundary correctly in many cases in which the human perceptual system, working only on acoustic-phonetic data, would probably make the assignment incorrectly. Another point to note is that for cases when T is greater than a certain limit (140 msec for voiced stop consonants and 200 msec for voiceless stop consonants) nothing can be said about the position of a word boundary (except that one has occurred!). Such long word boundaries are generally associated with important syntactic or emphatic events.

6.9 COARTICULATION

In Chapter 5 it was noted that the formant positions of the vowels in the VCV sequences considered in the experiment described in Section 6.4 did not change significantly as a result of coarticulation effects. Thus, in this section only coarticulatory effects on the plosive consonants are investigated.

There are three types of parameters associated with plosive consonants where coarticulatory effects could possibly be manifested. These are timing parameters, and parameters related to formant transitions and burst spectra. Each is discussed in turn.

6.9.1 Coarticulation and timing parameters

One timing parameter which could show some coarticulatory variation is T the time interval between the endpoints of the vowel transitions (defined in Section 6.6.3). It has already been shown that T depends on place of articulation, voicing manner, and juncture position. But as was already explained it incorporates an indirect measure of VOT and Klatt (1975) has shown that VOT varies somewhat according to the identity of the vowel following the consonant.

It was, however, hard to see any systematic coarticulatory variations in T for this set of data.

Another timing parameter which could show coarticulatory effects is the time associated with the formant transitions from the steady state of the vowel to or from the plosive consonant. One problem with this parameter is that it is difficult in many cases to decide exactly where the transition begins. Speaker 10, for example, produces vowels which have practically no steady state; the vowel formants are gradually changing all the time. The other speakers, however, generally have a well-defined vowel steady-state region followed by transitions.

In general coarticulatory effects on formant transition lengths are not strong. The most noticeable effect was that the length of transition from the vowel /i/ to the consonant /d/ is generally significantly shorter than the length of transitions from the vowel /ɔ/ to the consonant /d/. (10-50 msec for /i/-/d/ transition; 20-80 msec for /ɔ/-/d/ transitions.) Some other noticeable effects included:

(1) The /b/-/i/ transition in all speakers' productions of /ɔ bi/ (from 'raw beef') is longer than all other /b/-vowel transitions for any

other context.

(2) Both the /i/-/d/ and the /d/-/i/ transitions of all speakers productions of /i di/ are much shorter than the average transition lengths for other vowel contexts of / /.

(3) The / /-/ / transitions in both /ɔt i/ and /ɔ ti/ are longer than / /-vowel transitions in other contexts for all speakers.

It should be noted here that transition lengths are also somewhat dependent on the place of articulation of the consonant. In particular it is very noticeable that transitions from vowels to the velar plosives (/k/, /g/) and from velar plosives to vowels are on average much shorter than transitions to and from labial and alveolar plosives.

6.9.2 Transition locus effects

The Locus Theory of formant transitions in stop consonants was discussed in Chapter 4 (Section 4.1). The Locus Theory (Delattre et al., 1955) predicts that formant transitions from the steady state of a vowel to a neighbouring stop consonant point to a particular frequency or locus depending on the place of articulation of the consonant. The locus values predicted were:

for /b/	720 Hz
for /d/	1800 Hz
for /g/ before front vowel	>1200 Hz
for / / before back vowel	not found

Ohman (1966) showed that in real speech the situation was somewhat more complex. He showed that for /b/ and /d/ the position to which the transitions pointed was dependent not only on the place of articulation of the consonant but also on the second formant of the vowel preceding the

consonant. Thus the second formant /b/ locus could be anywhere in the range 500-1400 Hz; being low if the second formant of the preceding vowel was low and high if the second formant vowel was high. A similar situation occurred for /d/ with the locus occurring in the range 1400-1700 Hz. For /g/ the formants pointed to a locus which was determined by the vowel following the /g/ if the vowel preceding the /g/ was a front or central vowel, and to a low locus if the vowel preceding the consonant was a back vowel.

The data examined in the experiment described in Section 6.1 revealed the following locus results (which can be seen in Figures 6.6(a)-(b)).

1. that in continuous speech (at least for Australian English) the position of the second formant locus is primarily determined for ALL consonants by the vowel preceding the consonant. This is more general than Ohman's result;

2. that the position of the F2 /b/ locus can range for male voices from 500 Hz (for the case when the preceding vowel is /ɔ/) to 1300 Hz for the case where the preceding vowel is /i/;

3. that the position of the F2 /d/ locus ranges for male voices from 1350 Hz for the case where the preceding vowel is /ɔ/ to 1750 Hz when the preceding vowel is /i/;

4. that there are high and low loci for /g/. If the preceding vowel is /i/ the /g/ locus for male voices is in the range 2000-2300 Hz. If the preceding vowel is /ɔ/ the /g/ locus for male voices is in the range 800-1500 Hz;

5. that, at least for the case where the consonant is /g/, the exact position of the locus is also determined by the vowel following the

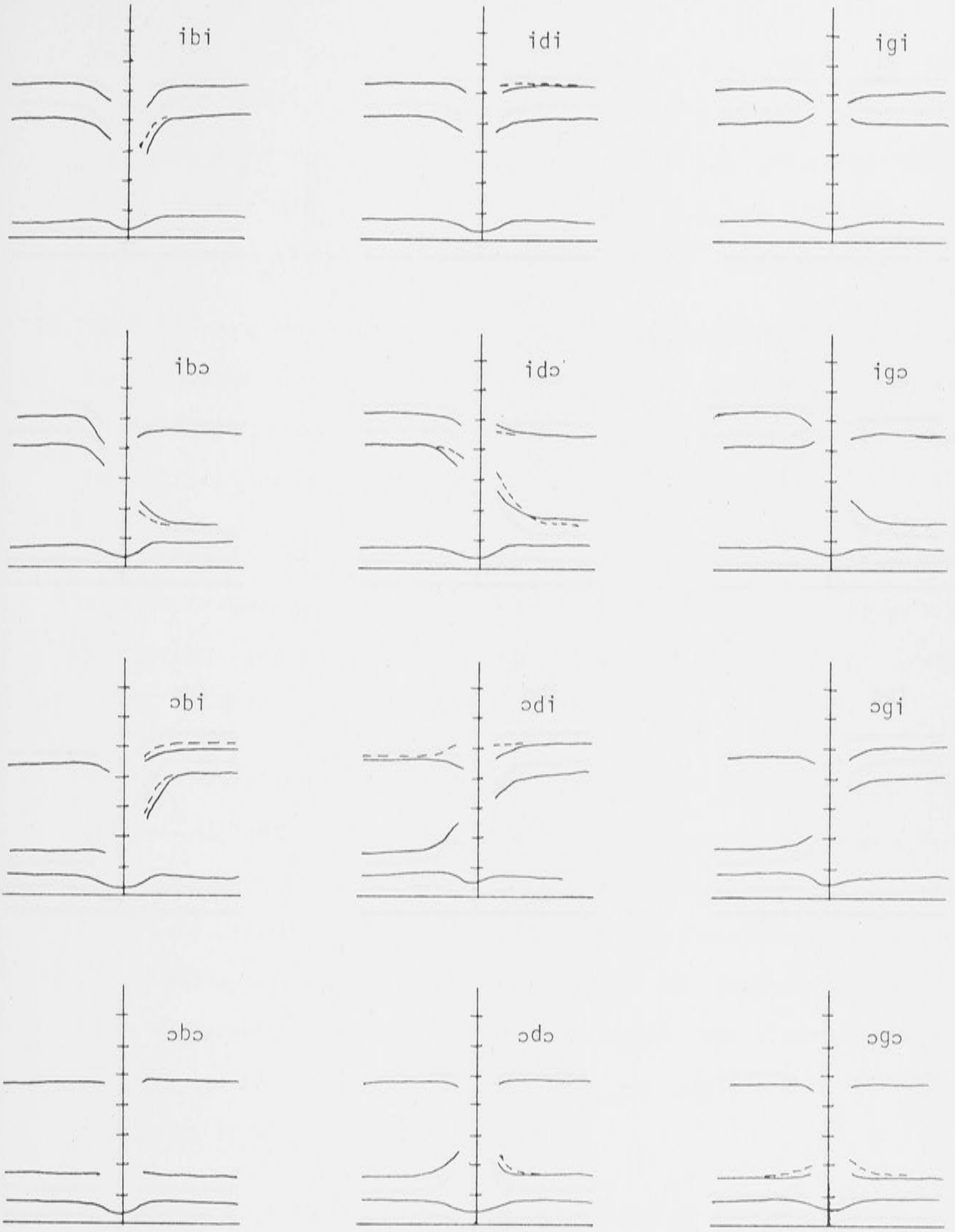


FIGURE 6.6(a): Average, first, second and third formants for male voices VCV sequences in which C is a lax plosive. Where dotted lines are shown they represent the /V CV/ case in contradistinction to the /VC V/ case (full lines). Where only a full line is shown both cases essentially coincide. Frequency marks are 0.5 kHz apart. Time is not strictly to scale but is of the order of 450 msec, in total.

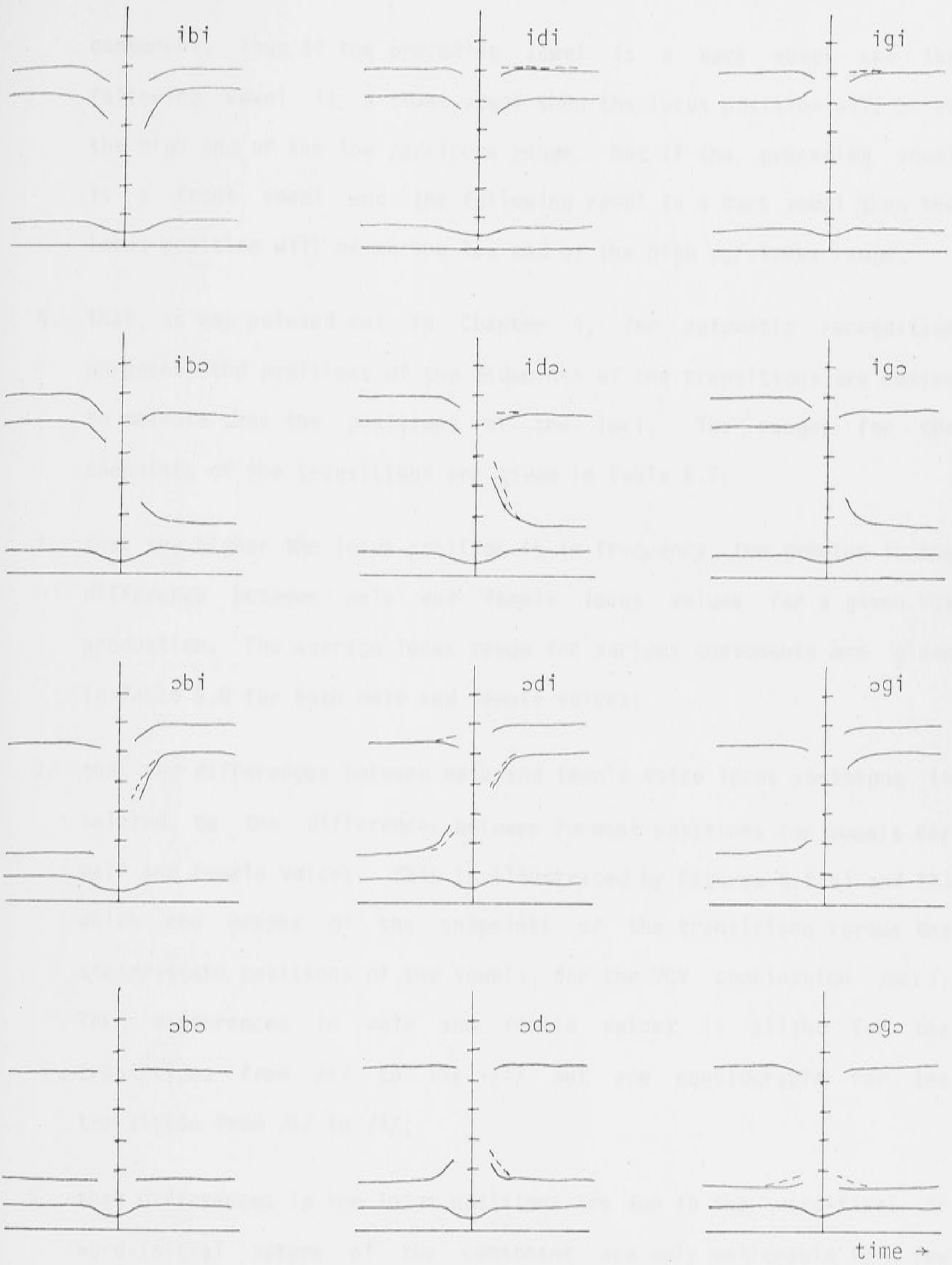


FIGURE 6.6(b): Average, first, second and third formants for female voices for VCV sequences in which C is a lax plosive. Where dotted lines are shown they represent the /V CV/ case in contradistinction to the /VC V/ case (full lines). Where only a full line is shown both cases essentially coincide. Frequency marks are 0.5 kHz apart. Time is not strictly to scale but is of the order of 450 msec, in total.

consonant. Thus if the preceding vowel is a back vowel and the following vowel is a front vowel then the locus position will be at the high end of the low /g/-locus range. But if the preceding vowel is a front vowel and the following vowel is a back vowel then the locus position will be at the low end of the high /g/-locus range;

6. that, as was pointed out in Chapter 4, for automatic recognition purposes the positions of the endpoints of the transitions are easier to measure than the positions of the loci. The ranges for the endpoints of the transitions are given in Table 6.7;

7. that the higher the locus position is in frequency, the greater is the difference between male and female locus values for a given VCV production. The average locus range for various consonants are given in Table 6.8 for both male and female voices;

8. that the differences between male and female voice locus positions is related to the differences between formant positions for vowels for male and female voices. This is illustrated by Figures 6.7(a) and (b) which are graphs of the endpoints of the transitions versus the steady-state positions of the vowels, for the VCV combination /ɒdi/. The differences in male and female voices is slight for the transitions from /ɒ/ to the /d/ but are considerable for the transition from /d/ to /i/;

9. that differences in the locus positions are due to the word-final or word-initial nature of the consonant are only noticeable in a few cases. If an effect is present at all it is often only noticeable in the CV transition. As can be seen in Figures 6.6(a) and (b) the most noticeable cases in which juncture effects are present are /ɒbi/, /idɒ/, /ɒdɒ/, /ɒg/;

	Labial Locus Range	Alveolar Locus Range	Low Velar Locus Range	High Velar Locus Range
Male Voices	500-1300	1350-1900	800-1500	2000-2300
Female Voices	500-1500	1550-2400	800-1800	2400-2900

TABLE 6.7: Locus ranges of plosive consonants at the three places of articulation for male and female voices.
All measurements in Hz.

	Speaker	Frequency Range of First Peak	Frequency Range of Second Peak
Male Speakers	1	0.5-2.6	3.6-4.3
	2	0.9-2.4	3.2-4.0
	3	0.8-2.7	3.4-4.4
	4	0.9-3.2	3.2-4.2
	5	0.8-2.7	3.1-4.4
Female Speakers	6	0.7-2.7	4.3-5.3
	7	0.7-3.6	3.9-4.8
	8	0.8-3.3	4.5-5.3
	9	1.2-3.4	3.6-5.2
	10	0.9-3.0	4.5-5.1

TABLE 6.8: Frequency ranges of velar burst peaks for the five male and five female speakers.
All measurements in kHz.

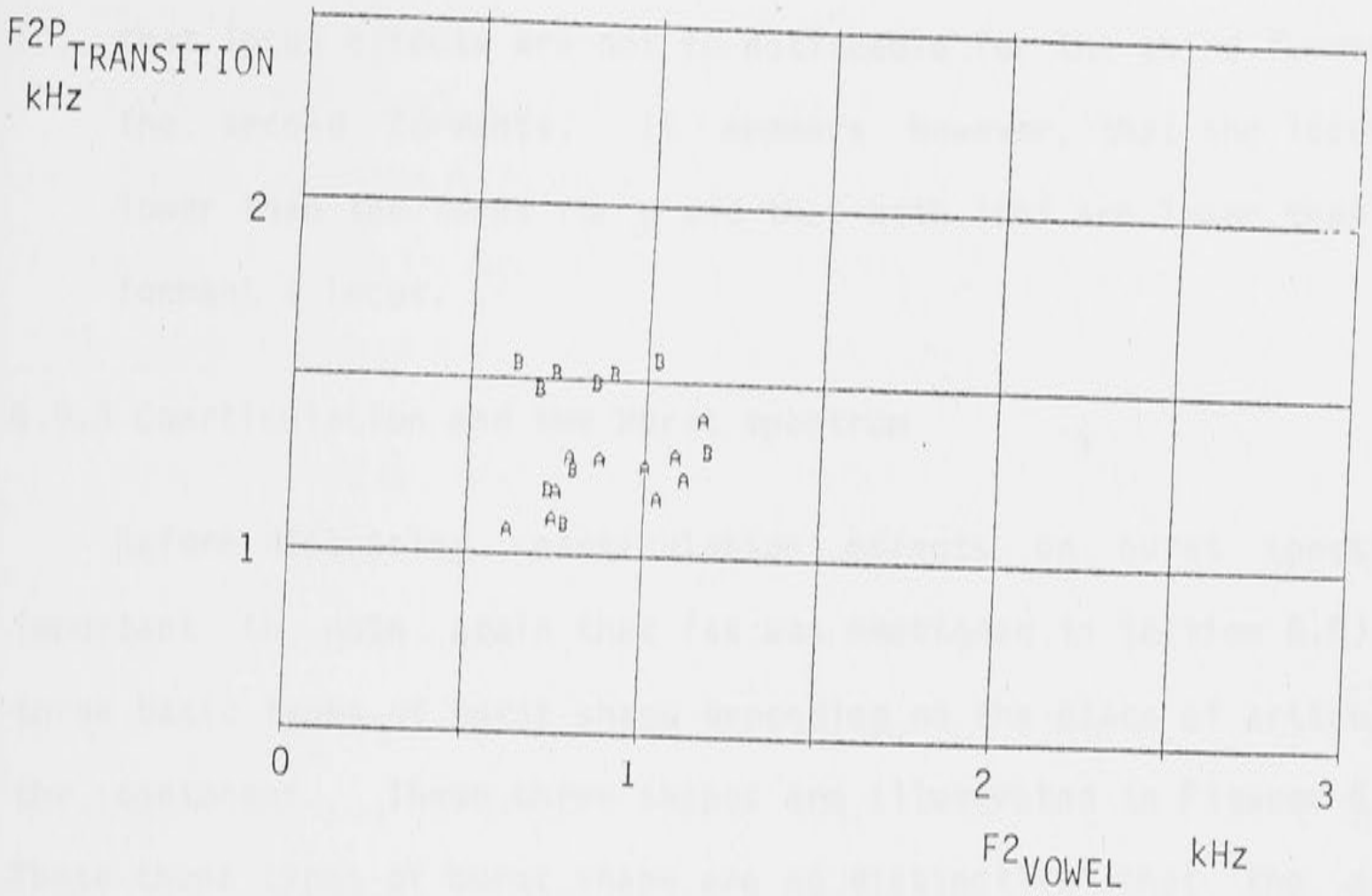


FIGURE 6.7(a): Endpoints of transitions from /ɔ/ to /d/ in /ɔdi/ versus $F2_{/ɔ/}$.

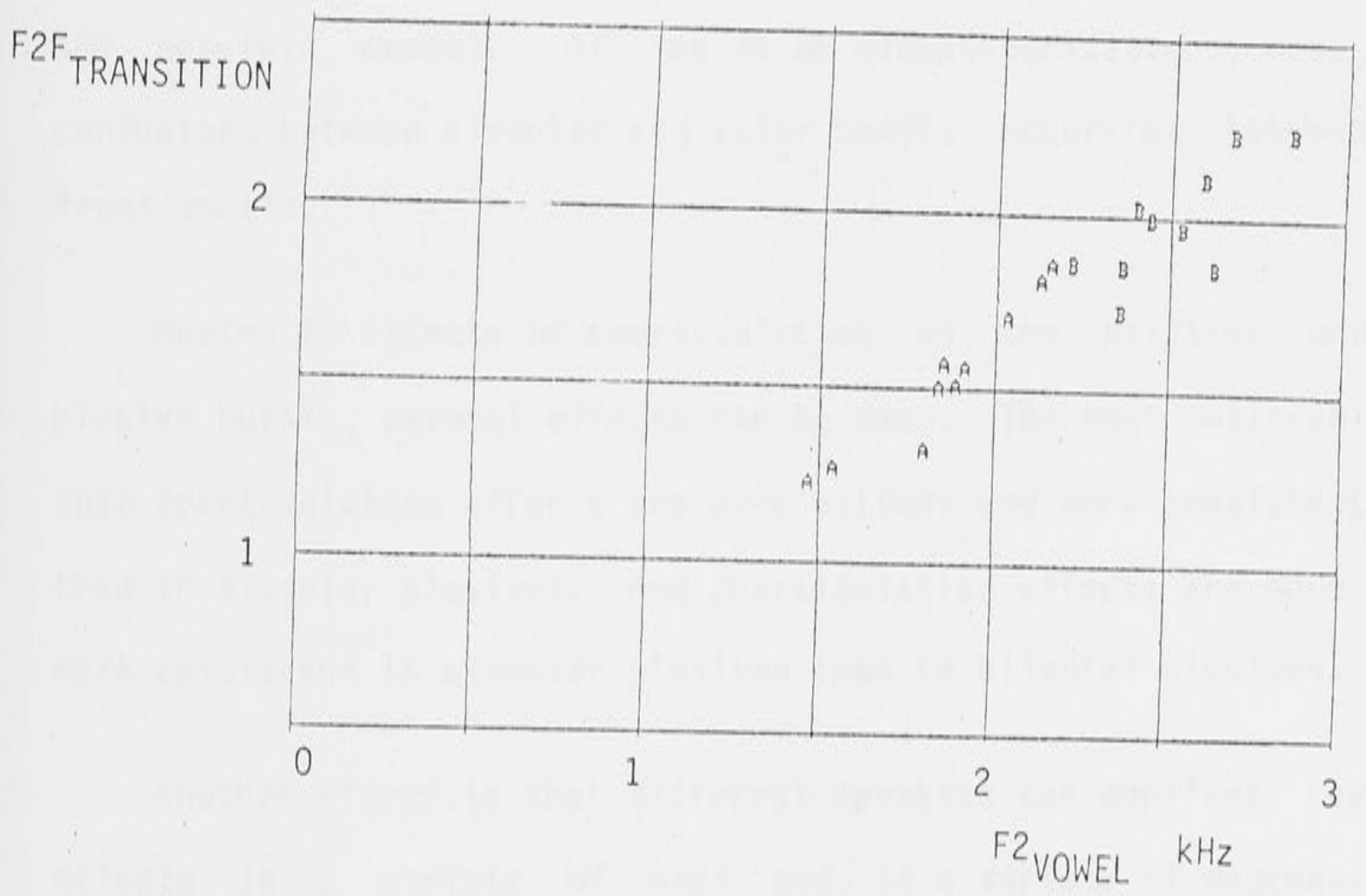


FIGURE 6.7(b): Endpoints of transitions from /d/ to /i/ in /ɔdi/ versus $F2_{/i/}$.

Note: A - male voices
B - female voices

10. that locus effects are not so noticeable for the third formants as for the second formants. It appears however, that the locus for b is lower than the locus for g and that both loci are lower than the third formant d locus.

6.9.3 Coarticulation and the burst spectrum

Before discussing coarticulation effects on burst spectra it is important to note again that (as was mentioned in Section 6.5) there are three basic types of burst shape depending on the place of articulation of the consonant. These three shapes are illustrated in Figures 6.4(a)-(c). These three types of burst shape are so distinctive that the author can correctly categorize by visual inspection 95% of the cases in the experimental corpus where bursts are produced (i.e. in 460 cases out of the 480 possible cases). Of the 5% of miscategorizations, most were due to confusions between alveolar and velar bursts occurring between two high front vowels.

Moving to effects of coarticulation on the position and shape of plosive bursts, several effects can be seen. The most noticeable effect is that coarticulation effects are more evident and more consistent in velar than in alveolar plosives. And coarticulation effects are more evident and more consistent in alveolar plosives than in bilabial plosives.

Another effect is that different speakers can manifest coarticulation effects in a variety of ways and to a variety of degrees. Thus only speakers 1, 2 and 10, show any coarticulation effects at all in bilabial bursts. Speakers 4 and 8 do not use any coarticulation effects in the production of alveolar bursts while all other speakers do. All speakers

use coarticulation effects in the production of velar bursts. However, while most speakers show strongest coarticulation effects with the vowel following the velar bursts, speaker 6 seems to show about equal coarticulation effects between the consonant and both the preceding and the following vowels. Some speakers, notably speakers 1, 2, 4 and 10 display interaction between coarticulation and juncture effects.

Coarticulation, when it occurs, in bilabial bursts is manifested by the burst 'hump' appearing above 2 kHz if the bilabial consonant occurs between two front vowels and appearing below 1.5 kHz if the consonant occurs between two back vowels. Of the speakers who do not display any systematic effects in bilabial burst production, speaker 7 is unusual in that she produces bilabial bursts which are highly non-coarticulatory in that all such burst spectra are approximately the same shape and occur in approximately the same position regardless of context.

Coarticulation effects in alveolar bursts are manifested by the burst for an alveolar consonant occurring between two high front vowels being reasonably diffuse with the burst 'hump' occurring at a frequency greater than 3.5 kHz, while the burst of an alveolar consonant occurring between two back vowels tends to be more compact in shape with the burst 'hump' occurring below 4 kHz.

Coarticulation effects in velar bursts are quite spectacular. Such effects are most strongly indicated by the position of the lower in frequency (and generally higher in amplitude) of the two peaks of the typical velar spectrum. The position of this first 'peak' for the case when the consonant occurs in the /i-i/ frame is high (about 2.7 kHz for male voices; 3.2 kHz for female voices). In this case the position of the second peak tends to be low (3.4 kHz for male voices; 3.9 kHz for female

voices) and often the two peaks superimpose to give the appearance of a single slightly diffuse peak. This is why it is often impossible to tell from the burst shape in the /i-i / frame whether the consonant was alveolar or velar. For bursts of consonants occurring in the /ɔ-i/ frame the position of the first peak is lower in frequency than it is for the /i-i/ frame but higher than it is for the /i-ɔ/ frame. For consonants occurring in the /ɔ-ɔ/ frame the position of the first peak is lower in frequency than it is for consonants occurring in any other context. The relative positions of velar burst peaks are illustrated in Figure 6.8 for speaker 3 (male) and speaker 7 (female). In summary both vowels affect the velar burst position, but the following vowel has a stronger effect than preceding vowels.

Although the positions of the first peaks of velar bursts occur in the sequence described above for the majority of cases, sometimes juncture and coarticulation effects interact such that for consonants in word-final positions there are two opposing effects operating:

(1) the coarticulation effects outlined in the previous paragraph;

(2) the tendency for a consonant to be attached to the vowel on the same side of the word boundary as the consonant.

This second effect is, not surprisingly, very noticeable when the word boundary is associated with a long time interval. The net effect of the two effects is that the position of the first peak of /ɔgi/ or /ɔki/ will be lower in frequency than the first peak of /igɔ/ or /ikɔ/ in some cases. Nevertheless it must again be stressed that, in general, burst coarticulation with the vowel following the consonant is the predominant coarticulation effect regardless of juncture position.

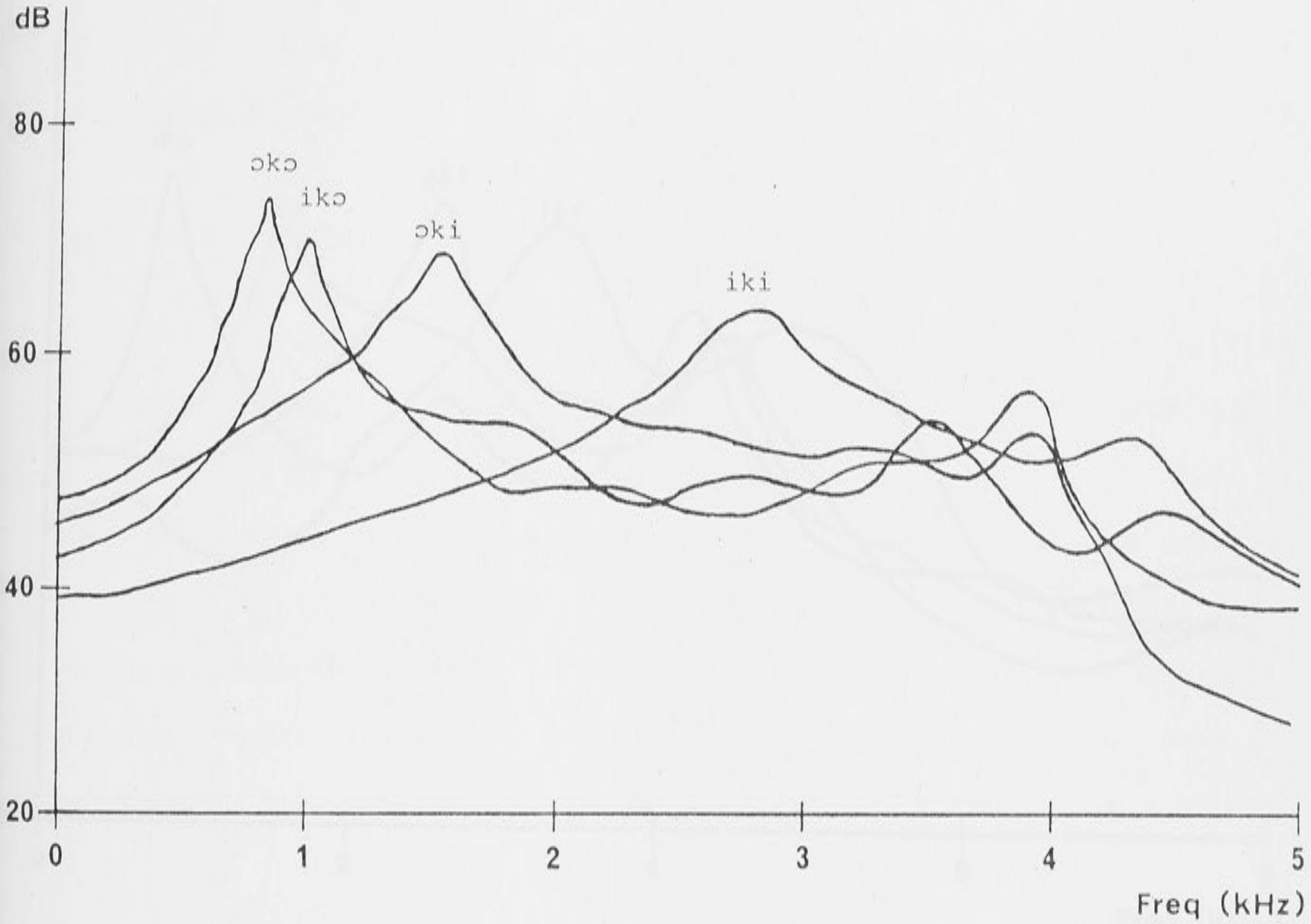


FIGURE 6.8(a): Velar bursts showing coarticulation effects for Speaker 3 (male).

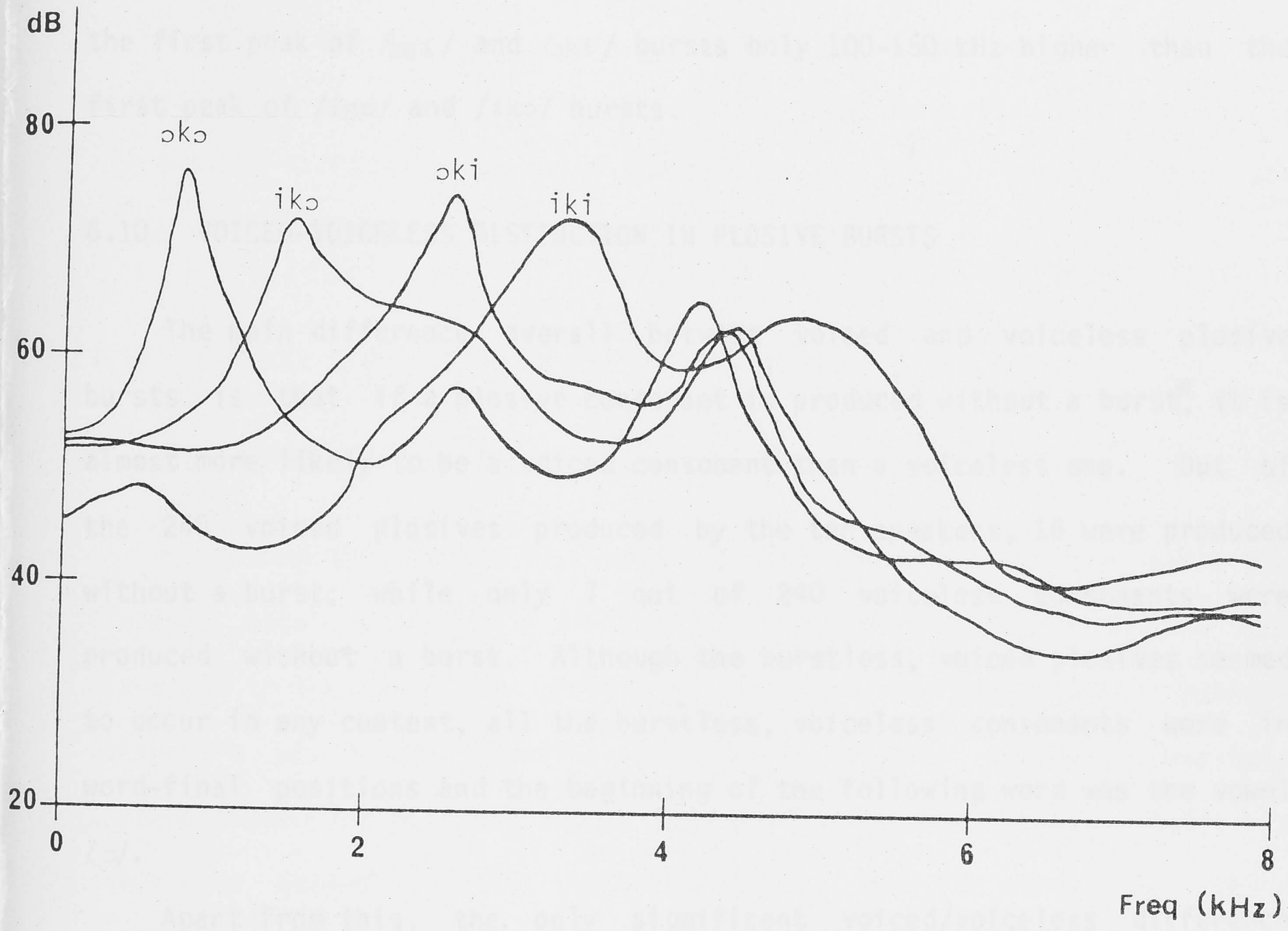


FIGURE 6.8(b): Velar bursts showing coarticulation effects for Speaker 7 (female).

Finally, it should be mentioned that the degree of coarticulation with the vowel following the consonant is not always very much greater than the degree of coarticulation with the vowel preceding the consonant. This is particularly so for speaker 6 who, regardless of the juncture position, has the first peak of /ɔgi/ and /ɔki/ bursts only 100-150 kHz higher than the first peak of /igɔ/ and /ikɔ/ bursts.

6.10 VOICED-VOICELESS DISTINCTION IN PLOSIVE BURSTS

The main difference overall between voiced and voiceless plosive bursts is that if a plosive consonant is produced without a burst*, it is almost more likely to be a voiced consonant than a voiceless one. Out of the 240 voiced plosives produced by the ten speakers, 16 were produced without a burst; while only 7 out of 240 voiceless consonants were produced without a burst. Although the burstless, voiced plosives seemed to occur in any context, all the burstless, voiceless consonants were in word-final positions and the beginning of the following word was the vowel /ɔ/.

Apart from this, the only significant voiced/voiceless difference occurs in alveolar bursts. For almost all of the ten speakers, /t/ burst 'humps' were more massive and the hump position was at a higher frequency than for /d/ bursts.

6.11 MALE-FEMALE DIFFERENCES IN PLOSIVE BURSTS

There are no male/female differences in bilabial bursts. For alveolar bursts the main concentration of energy (the burst 'hump') occurs anywhere in the range 1.5-5 kHz for male voices; but extends to higher frequencies for female voices where the range is 1.5-6.5 kHz. For velar bursts, the

* 'Without a burst' means that no burst could be detected.

ranges of the two peaks for various speakers are given in Table 6.9. From this table it can be seen that the second peak very strongly reflects male/female differences. The position of the first peak reflects male/female differences only in some contexts. For velar plosives in the /i-i/ context the differences are marked; they are progressively less marked for velar plosives in /ɔ-i/ and /i-ɔ/ contexts; and there is no discernable difference for velar plosives in the /ɔ-ɔ/ context when the first peak is always below 1 kHz.

6.12 SPEAKER DIFFERENCES IN PLOSIVE CONSONANTS

6.12.1 - Male/female differences

Combining the results discussed in the previous section with the findings about male/female differences in vowel-consonant transition loci discussed in Section 6.9.2 (Point 8), it can be seen that consonant parameters which are measures of absolute frequency, become increasingly differentiated for male and female speakers as the frequency region in which the parameter falls becomes higher. Thus for measurements below 1 kHz there is no difference according to the sex of the speaker while for measurements above 2 kHz the differences between male and female voices will be quite marked. The actual difference between the average male and female voices for consonant parameters that are measures of absolute frequency is given by the curves in Figure 5.5 which gives the positions of the second formant of a 'continuum' of vowels for both male and female speakers. Thus it seems reasonable to predict that absolute frequency parameters will display male/female differences in a similar fashion for all classes of sounds of Australian English. This is important as it allows one to devise automatic recognition algorithms for female voices

from algorithms that have been devised (as most algorithms have) from measurements on male voices only.

Furthermore it is important to note that automatic recognition algorithms which attempt to circumvent the problems of male/female differences by scaling male frequency measurements by a constant percentage (often 20%) to obtain the corresponding female values will, in general, give incorrect results because of the non-linear (in frequency) relationship between male and female voice parameters.

In the preceding discussion the emphasis has been on the average male and female voices. As is obvious from many of the tables and figures which have been used to illustrate the systematic nature of the differences between the the ten speakers in the data set, in actual fact there are generally not two distinct regions - one male, one female - in frequency space but rather there is a 'continuum' with some male voices in the 'lower' end of the female frequency range, and some female voices the 'higher' end of the male frequency range for some particular parameter. particular parameter. In the recognition algorithms presented in Chapter 5 and in this chapter, the notion of the <u>average</u> male voice and the <u>average</u> female voice is not used. Rather the algorithms check to see a parameter falls into the appropriate <u>range</u> for either a male or a female voice producing the sound hypothesised.

6.12.2 - Speaker idiosyncracies

Throughout the discussion of the experiment described in Section 6.4, speaker idiosyncracies and wide variations between speakers have been noted. The amount of variation between speakers in the production of a certain sound depends on the identity of the sound. For example, coarticulation patterns associated with velar consonants are very similar

for all speakers while in the production of alveolar consonants some speakers will display significant coarticulation; others, none at all.

The important implication for automatic recognition algorithm design is that, recognition procedures will be more likely to be applicable to a wide range of speakers if they depend on parameters which display minimal speaker dependence (this applies both along the male/female axis and along the speaker idiosyncratic scale). Where it is necessary to use parameters which might be speaker dependent, suitable allowance must be made for possible variations in productions between speakers. Speaker dependence theoretically should be predictable from studies of the production mechanism of various sounds. At least, it is generally easy in retrospect to find a production rationale for speaker variations that have been discovered accidentally!

6.13 RECOGNITION RULES FOR PLOSIVE CONSONANTS IN AUSTRALIAN ENGLISH

6.13.1 - Overview

For the recognition of plosive consonants in Australian English, as in the recognition of plosive consonants in Italian, two types of acoustic features are considered, namely formant transitions (FOR) and burst spectra (SP). As in Chapter 4, the degree of compatibility between $p\{t_i, t_j\}$, an acoustic pattern in the time interval (t_i, t_j) which has been previously evaluated as being both 'sonorant' and 'interrupted', and X, a generic plosive consonant, can be written as the following fuzzy semantic rule:

$$\begin{aligned} \text{Poss}(X \text{ is in } p) = \; & \alpha_X \wedge \text{Poss}(XFOR \text{ is in } p) \\ & \vee \beta_X \wedge \text{Poss}(XSP \text{ is in } p) \\ & \vee \gamma_X \wedge \text{Poss}(XFOR \text{ is in } p) \wedge \text{Poss}(XSP \text{ is in } p) \end{aligned}$$

which may be expressed in algebraic form as follows (by replacing max and min operations by multiplication and addition signs, respectively):

$$X = \alpha_X XFOR + \beta_X XFOR + \gamma_X XFOR.XSP$$

where α_X, β_X, γ_X [0.1] represent the 'grammaticalities' of the syntactic rules relating the two basic acoustic features (XFOR + XSP) and the combination of these features to the plosive X.

Again as in Chapter 4, the syntactic category, XFOR, is expressed in terms of the following features*:

XFL = formant pseudo-loci

XFS = formant slopes

which are further qualified by adding the letters (P) or (F) depending on whether the formant transition referred to precedes or follows the consonant. The formant rules will be considered in detail in the next section.

The rules for the burst spectra are quite different from the rules for burst spectra used in Chapter 4. The motivation for the type of burst rules considered here comes from the Stevens and Blumstein (1979) templates and will be discussed in Section 6.13.3.

Although rules for both the set P_L of lax plosives /b,d,g/ and the set P_T of tense plosives /p,t,k/ were generated, the similarities according to place of articulation /b ↔ p, d ↔ t and g ↔ k/ between the two sets are marked; therefore the rules for the lax plosives only will be given, and any significant differences between these rules and those for the tense

* In this set of rules, unlike the rules for Italian plosives, the buzz-bar characteristics were not found to be necessary.

plosives will be noted.

The rules were devised from consideration of the VCV utterances for eight speakers (Speakers 1-4 (male) and Speakers 6-9 (female)). These VCV utterances were obtained from the experiment described in Section 6.4. The rules were tested on the complete set of experimental VCV utterances from Speakers 5 and 10; on further VCV, VC and CV utterances from the ten speakers who participated in the experiment; and on random VCV utterances from several other adult speakers.

6.13.2 - Recognition rules relating to formant transitions

The two important features (XFL and XFS) which are used in the rules for formant transitions will be discussed in sequence.

The positions of the pseudo-loci of the second and third formants are just the positions of the 'endpoints of the transitions'. The method of locating these endpoints was described in Section 6.6.3. It should be noted that the definition for the endpoints of transitions preceding the consonant (XFLP) and the definition of endpoints of transitions following the consonant (XFLF) are not symmetrical.

The fuzzy restrictions defined over the pseudo-loci are characterised by vectors of breakpoints. The allowable positions of a pseudo-locus (whether relating to second or third formants; to preceding or following transitions; or to male or female voices) are expressed by a vector of subscripted symbols, belonging to the terminal alphabet V , which gives the labels of the fuzzy restrictions which are defined by a vector of numerical breakpoints. The vectors of symbols and their corresponding vectors of breakpoints for all the pseudo-loci considered by the recognition rules are given in Table 6.9.

VARIABLE	LABELS OF FUZZY RESTRICTIONS ON VARIABLE	VECTOR OF BREAK-POINTS ASSOCIATED WITH FUZZY RESTRICTIONS
F2 pseudo-loci male voices	$\{l_1, a_1, b_1, c_1, d_1, e_1, h_1\}$	{400,600,1000,1300,1500,1800,2100,2300}
F2 pseudo-loci female voices	$\{l_2, a_2, b_2, h_2\}$	{1800,2000,2300,2500,2800}
F2 pseudo-loci female voices	$\{l_3, a_3, b_3, c_3, d_3, e_3, f_3, g_3, h_3\}$	{400,600,1000,1350,1600,1900,2100,2400,2800}
F3 pseudo-loci female voices	$\{l_4, a_4, b_4, c_4, d_4, h_4\}$	{2200,2400,2600,2800,3000,3200,3500}

TABLE 6.9: Fuzzy restrictions and associated vector of break-points for pseudo-loci variables

Figures 6.9 (a) and (b) and 6.10 (a) and (b) display the fields of existence of pseudo-loci in F2-F3 space. Figure 6.9(a) displays, for male voices, the positions of the pseudo-loci preceding the consonant for the three classes of plosives in the various contexts given in square brackets. For simplicity, only the positions of /b,d,g/ are shown. Although the vowels used in the training set of utterances were only front vowel /i/ and the back vowel/ɔ/, the vowel contexts listed on the pseudo-loci diagrams are given as either front (f) or back (b) as it was found in tests that the rules applied almost as well to other front and back vowels. Central vowel context rules have not as yet been added to the rule set. Figure 6.9(b) displays the positions of pseudo-loci for transitions following the consonant for male voices. Figures 6.9 (a) and (b) correspond to Figures 4.1 (a) and (b) which give similar results for Italian plosives.

Figures 6.10 (a) and (b) are the figures for female voices corresponding to Figures 6.9 (a) and (b) for male voices. Although the two sets of figures show similar overall patterns, the pseudo-loci position

figures for female voices are not simple translations of the corresponding figures for male voices because of the non-linear relationship in frequency between corresponding features for male and female voices.

The recognition rules relating to pseudo-loci for male voices are:

$$\text{BFLP} = 1_2([f,f](c_1+d_1) + ([b,b](a_1+b_1)) + (a_2+b_2)([b,*](c_1+d_1)+[b,*](a_1+b_1)+[b,f]1_1)$$

$$\text{DFLP} = [f,*](d_1+e_1)b_2+(a_2+b_2)([b,*]b_1+[b,b]c_1)$$

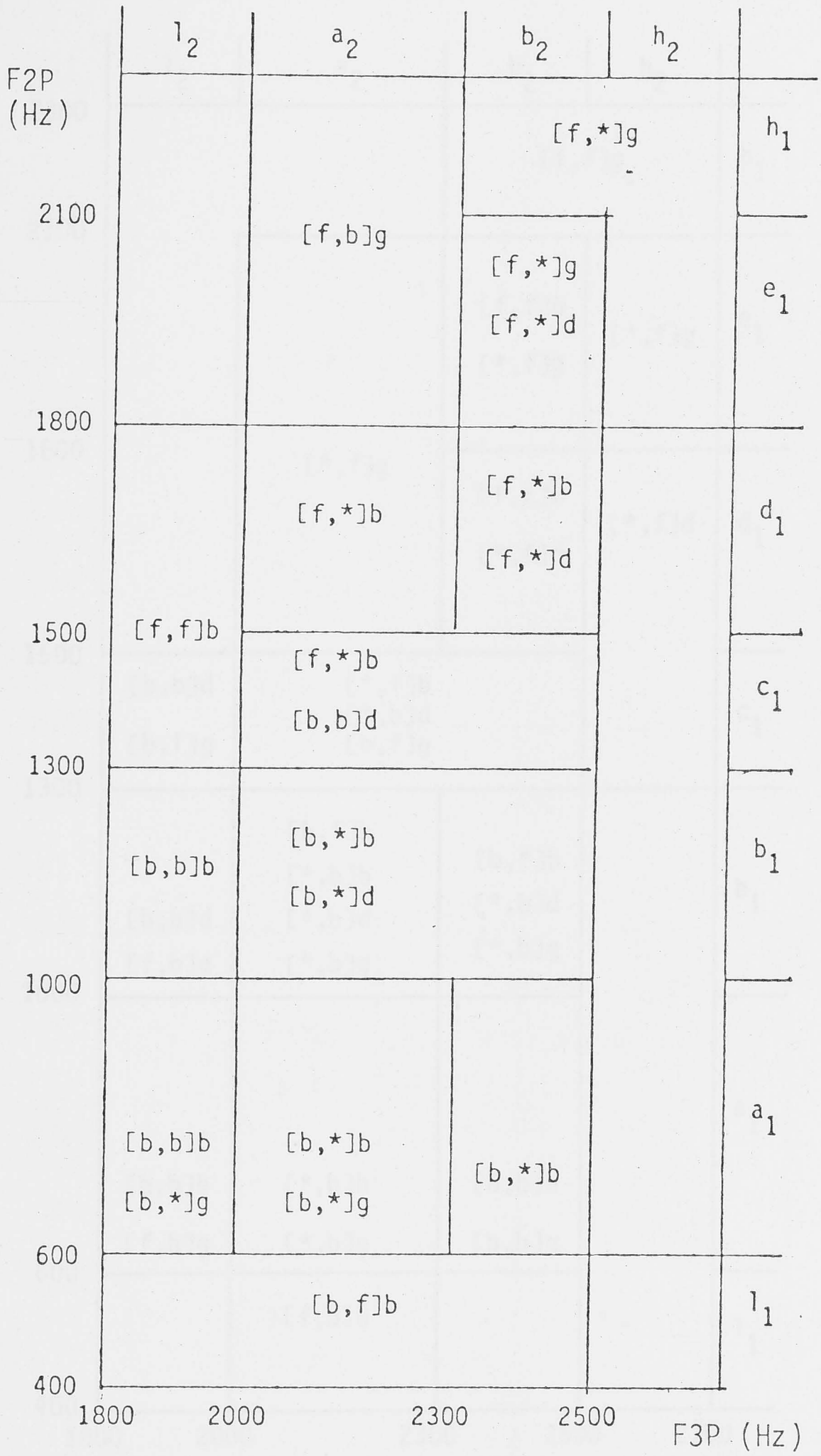


FIGURE 6.9(a): Fields of existence in F2-F3 space for male voices of the endpoints of transitions from a vowel to the plosive consonant following it. The contents of square brackets indicate the vocalic context. Fuzzy interval labels are marked along the top and right side of the figure.

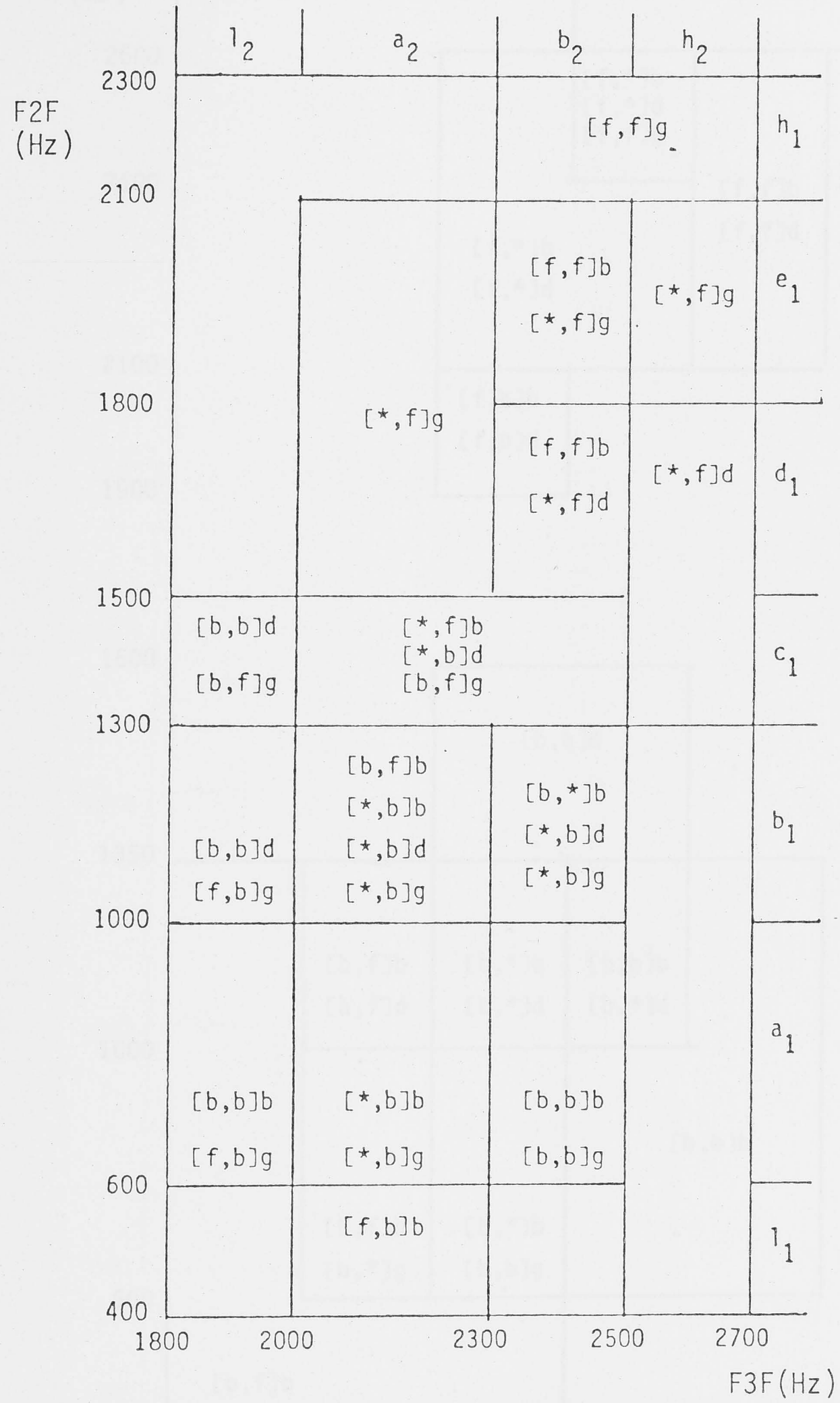


FIGURE 6.9(b): Fields of existence in F2-F3 space for male voices of the beginning-points of transitions from a plosive consonant to the vowel following it. The contents of the square brackets indicate the vocalic context. Fuzzy interval labels are marked along the top and right side of the figure.

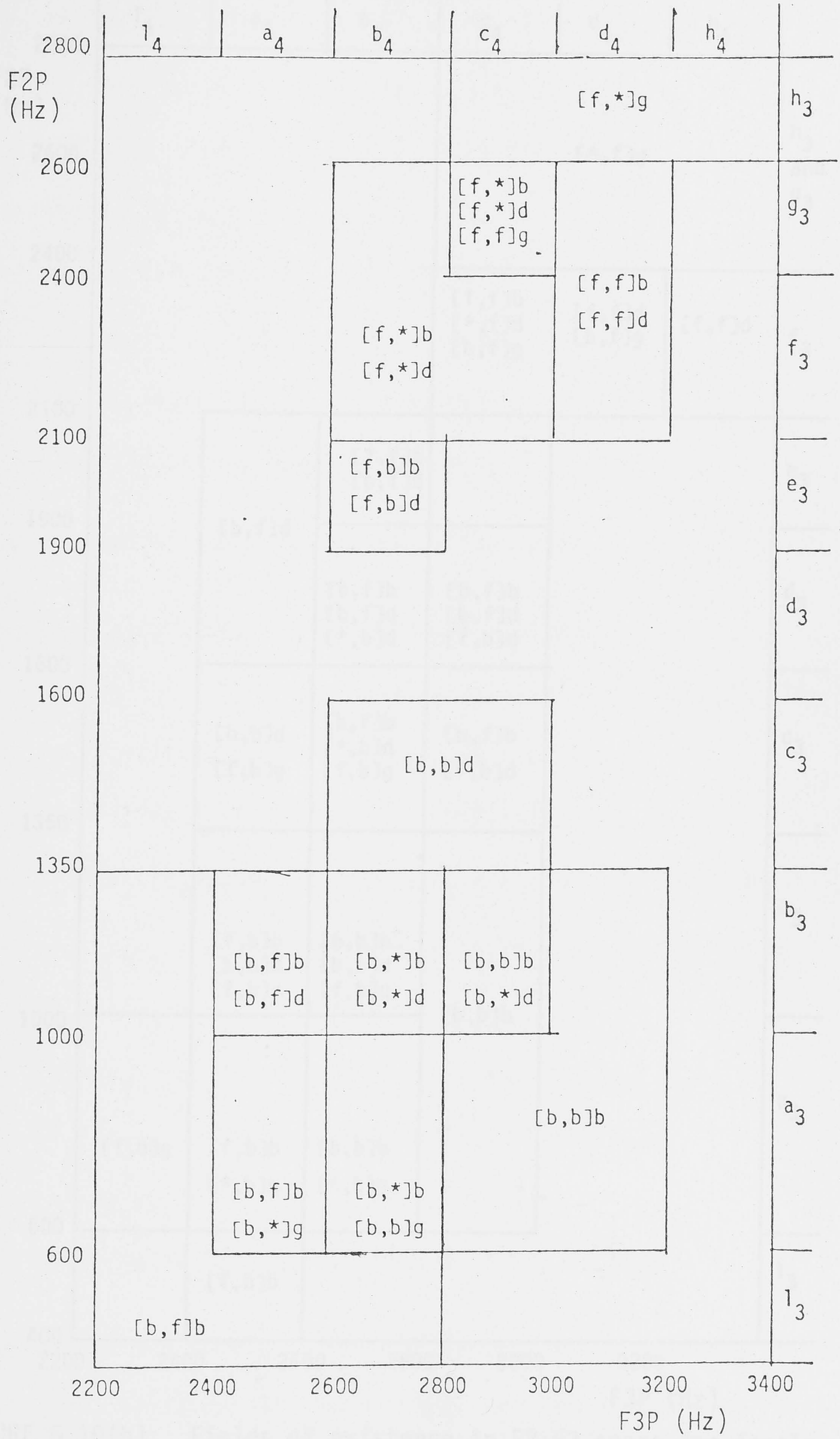


FIGURE 6.10(a): Fields of existence in F2-F3 space for female voices of the endpoints of transitions from a vowel to the plosive consonant following it. The contents of square brackets indicate the vocalic context. Fuzzy interval labels are marked along the top and right side of the figure.

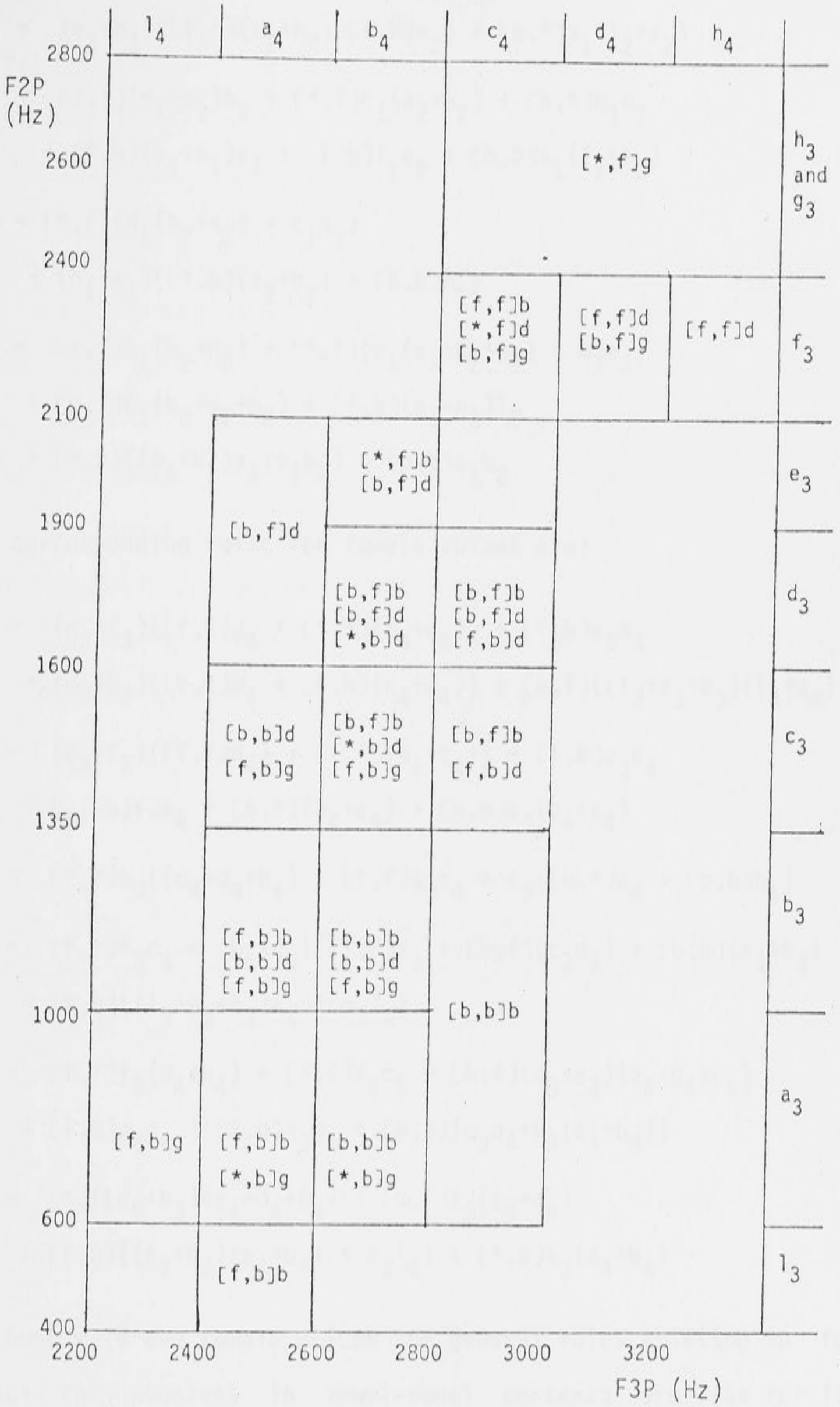


FIGURE 6.10(b): Fields of existence in F2-F3 space for female voices of the beginning-points of transitions from a plosive consonant to the vowel following it. The contents of the square brackets indicate the vocalic context. Fuzzy interval labels are marked along the top and right side of the figure.

$$GFLP = (e_1+h_1)([f,*](b_2+h_2)+[f,b]a_2) + [b,*]a_1(l_2+a_2)$$

$$BFLF = [f,f](d_1+e_1)b_2 + [*,f]c_1(a_2+b_2) + [b,*]b_1b_2$$
$$+ [*,b](a_1+b_1)a_2 + [f,b]l_1a_2 + [b,b]a_1(l_2+b_2)$$

$$DFLF = [*,f](d_1(b_2+h_2) + c_1h_2)$$
$$+ (b_1+c_1)([*,b](a_2+b_2) + [b,b]h_2)$$

$$GFLG = [f,f]h_1(b_2+h_2) + [*,f](e_1(a_2+b_2+h_2) + d_1a_2)$$
$$+ [b,f]c_1(e_2+a_2+b_2) + [f,b](a_1+b_1)l_2$$
$$+ [*,b]((a_1+b_1)a_2+b_1b_2) + [b,b]a_1b_2$$

The corresponding rules for female voices are:

$$BFLF = (g_3+f_3)([f,f]d_4 + [f,*](b_4+c_4)) + [f,b]e_3b_4$$
$$+ (a_3+b_3)([b,*]b_4 + [b,b](c_4+d_4)) + [b,f]((l_3+a_3+b_3)(l_4+a_4)+(l_3b_4)$$

$$DFLP = (g_3+f_3)([f,f]d_4) + [f,*](b_4+c_4)) + [f,b]e_3b_4$$
$$+ b_3([b,f]a_4 + [b,*](b_4+c_4) + [b,b]c_3(b_4+c_4)$$

$$GFLP = [f,*]h_3((c_4+d_4+h_4) + [f,f]g_3c_4 + a_3([b,*]a_4 + [b,b]b_4)$$

$$BFLF = [f,f]f_3c_4 + (b_4+c_4)([*,f]e_3 + [b,f](c_3d_3) + [b,b](a_3+b_3)$$
$$+ [f,b]((l_3+a_3+b_3)a_4 + a_3l_4)$$

$$DFLF = [f,f]f_3(d_4+h_4) + [*,f]f_3c_4 + [b,f](d_3+e_3)(a_4+b_4+c_4)$$
$$+ [f,b]c_3c_4 + [*,b]c_3b_4 + [b,b](c_3a_4+b_3(a_4+b_4))$$

$$GFLF = [*,f](g_3+h_3)(c_4+d_4+h_4) + [b,f]f_3(c_4+d_4)$$
$$+ [f,b]((c_3+b_3)(a_4+b_4) + a_3l_4) + [*,b]a_3(a_4+b_4)$$

For both male and female voices the general rules relating to formant pseudo-loci for plosives in vowel-vowel contexts are (as for Italian

plosives:

BFL = BFLP.BFLF

DFL = DFLP.DFLF

GFL = GFLP.GFLF

The formant slope, XFS, is again defined as the difference between the steady state formant-value of the vowel and the endpoint of the transition from the vowel towards the consonant. Four formant slopes are considered by the rules - the slopes associated with the second and third formant transitions preceding the consonant and the slopes associated with the second and third formants following the consonants. Subscripted labels A (ascending) and D (descending) are used to indicate whether the transition points upwards or downwards into the consonant. The vector of labels and the vector of breakpoints for all the formant slopes considered are as follows

$[A_4, A_3, A_2, A_1, D_1, D_2, D_3, D_4]$

and

[1333, 1000, 667, 333, 0, -333, -667, -1000, -1333]

The rules for the formant slopes, XFSNY (where the N is the formant number, and the Y is P (preceding) or F (following)) are:

BFS2P = $[f,*](D_1+D_2) + [b,*](A_1+D_1)$

DFS2P = $[f,f]D_1 + [f,b](D_1+D_2) + [b,*](A_1+A_2) + [b,b]A_3$

GFS2P = $[f,*]A_1 + [f,f]A_2 + [b,*]A_1$

BFS2F = $[f,f](D_1+D_2) + [b,f](D_2+D_3+D_4) + [f,b](A_1+A_2+A_3)$
$+ [b,b](A_1+A_2+D_1)$

DFS2F = [f,f](D_1+D_2) + [b,f]D_2 + [*,b](A_1+A_2+A_3+A_4)

GFS2F = [*,f](A_1+D_1) + [f,b](A_1+A_2) + [b,b](A_1+D_1)

BFS3P = [f,*](D_1+D_2) + [b,*](A_1+D_1)

DFS3P = [f,*](D_1+D_2) + [b,*](A_1+A_2+A_3)

GFS3P = [f,f](A_1+A_2) + [f,b]A_1 + [b,f](A_1+A_2) + [b,b]A_1

BFS3F = [f,f](D_1+D_2) + [b,f](D_2+D_3+D_4) + [f,b](A_1+A_2+A_3) +
[b,b](A_1+A_2+D_1)

DFS3F = [*,f](D_1+D_2) + [*,b](A_1+A_2+A_3+A_4)

GFS3F = [f,f](A_1+D_1) + [b,f](A_1+D_1+D_2) + [f,b](A_1+A_2) + [b,b]A_1

For plosives in vowel-vowel contexts the most commonly used rules relating to formant slopes are:

BFS = BFS2P.BFS2F.BFS3P.BFS3F

DFS = DFS2P.DFS2F.DFS3P.DFS3F

GFS = GFS2P.GFS2F.GFS3P.GFS3F

It should be noted that the formant slope rules apply equally well to both male and female voices.

The pseudo-loci and formant slope rules are combined to give the following recognition rules for plosive formant transitions, XFL, assuming the plosive occurs between two vowels

$$BF = BFL(x).BFS + 0.5BFLP(x).BFSP + 0.5BFLF(x).BFSF$$

$$DF = DFL(x).DFS + 0.5DFLP(x).DFSP + 0.5DFLF(x).DFSF$$

$$GF = GFL(x).GFS + 0.5GFLP(x).GFSP + 0.5GFLF(x).GFSF$$

The symbols associated with the formant loci are written as a function of x to indicate that the particular rules used depend on whether the speaker is male or female.

The first term in each of the rules would be used when both second and third formant transitions from the surrounding vowels towards the consonant are well defined. The second and third terms would be used when for some reason, such as a long pause, one set of transitions was not well defined. The only major difference between the formant rules for Italian plosives and this set of formant rules, besides the omission here of the buzz-bar rules, is the lack in this set of a rule relating solely to formant loci. It was found for Australian English that correct and unique recognition was more likely if all formant rules depended on both formant loci and formant slopes.

6.13.3 - Rules relating to plosive bursts

Although the burst is located in a manner similar to that used in Chapter 4, the method of classifying plosive bursts in Australian English is quite different to that used for classifying plosive bursts in Italian.

After the burst has been located, the frequency spectrum at this point is considered. This frequency spectrum is calculated from the reflection coefficients (obtained from autocorrelation linear prediction analysis) and

is just the gain divided by the increase filter polynomial. Examples of such frequency spectra are given in Figures 6.4.

Three basic template shapes are used to classify the three different types of bursts - bilabial, alveolar, and velar. The labial bursts are characterised by a diffuse-falling template* which is displayed in Figure 6.11(a). For a burst to be classified as a /b/ burst with fuzzy membership 1.0, the outline of the frequency spectrum must not cross the two inner lines of the template and indeed must go between them. If the burst outline crosses the inner full line but not the dotted line it will be classified as a /b/ burst with membership 0.5.

The diffuse-falling template shown in Figure 6.11(b) is used to classify /b/ bursts in both male and female speech. /p/ bursts are classified by a diffuse-falling template that is slightly wider than the /b/ burst template but is the same general shape. Both diffuse-falling templates are coarticulation-independent, i.e. they do not change in shape for different vowel contexts surrounding the plosive consonant.

The bursts associated with the velar consonants, /g/ and /k/, are classified by a compact template which is illustrated in Figure 6.11(c). For a burst to be classified as a velar burst with membership 1.0 it must have two, narrow-bandwidth peaks which fall within the full-line 'triangles' in the template, and there must be at most one peak between these two peaks. Outlines of peaks which are inside the outer triangles (formed with the dotted lines) but not the inner triangles would be classified as velar bursts with 0.5 membership.

* The Blumstein and Stevens (1979) terminology is used here although the templates are defined slightly differently.

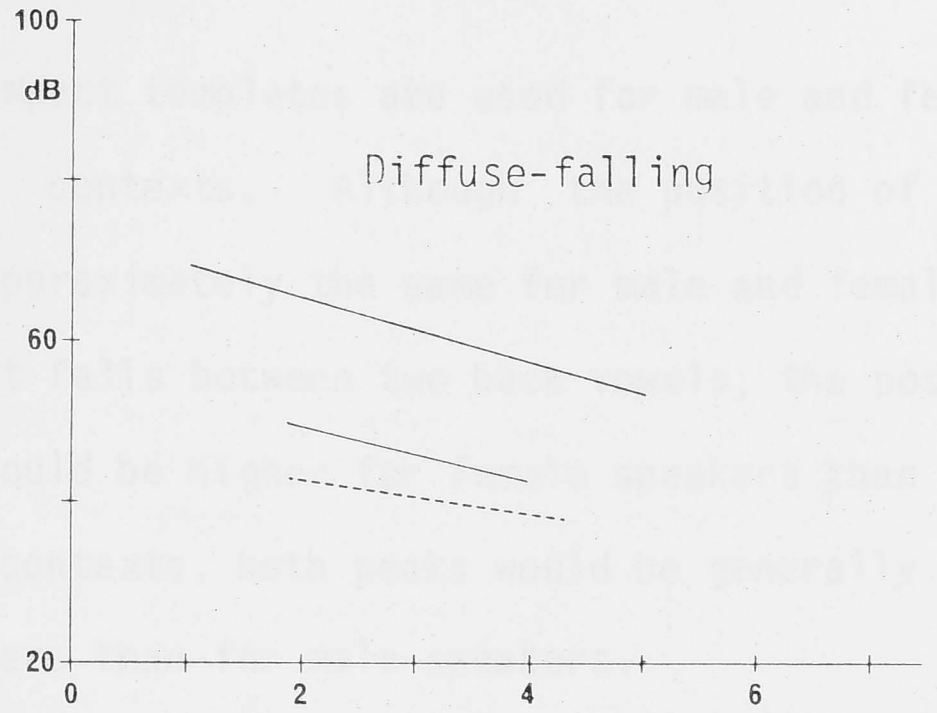


FIGURE 6.11: Diffuse-falling, diffuse-rising and compact templates used in the recognition of bilabial, alveolar, and velar consonant bursts, respectively.

Separate compact templates are used for male and female speech and for different vowel contexts. Although the position of the lower frequency peak should be approximately the same for male and female speakers when the plosive consonant falls between two back vowels, the position of the higher frequency peak would be higher for female speakers than for male speakers. For other vowel contexts, both peaks would be generally higher in frequency for female speakers than for male speakers.

The position of the triangle which captures the higher frequency peak varies only slightly with vowel context; the lower frequency peak, however, varies a good deal with context. The triangle template which captures this lower frequency peak is narrow and low in frequency if the consonant falls between two back vowels; it is wider and higher in frequency if one of the vowels on either side of the consonant is a back vowel and the other a front vowel; and is still wider and higher in frequency if the consonant occurs between two front vowels. The template for velar consonants occurring between two high, front vowels also allows for the possibility of the two characteristic peaks of the velar spectrum 'running-together' to form one rather wideband peak.

The alveolar consonants are characterised by a diffuse-rising template which is illustrated in Figure 6.11(b). For a plosive burst to be recognised as an alveolar burst, the burst spectrum outline must lie within the full lines of the diffuse-rising template. As usual, if it crosses the higher full line but falls within the dotted template lines it will be classified as alveolar with 0.5 membership, although one narrow bandwidth peak situated in the alveolar locus range (1400-2100 kHz) is allowed to cross the template lines without the burst losing its alveolar classification.

However, not only must a burst conform to the diffuse-rising template shape to be finally classified as alveolar, but it must also not be captured by the diffuse-falling or the compact templates. There is a slight problem with this requirement in that in a small number of cases the compact template for consonants occurring between two front vowels captures alveolar bursts in the same context. For this reason the velar and alveolar burst rules for this context do not have very high grammatical weightings in the final rules. By some sort of natural compensation, for the case of consonants occurring between front vowels the velar vowel-consonant transitions form a very distinctive pattern which is quite different to the pattern formed by transitions associated with other consonants. Also labial and alveolar consonants are well separated in this case by the large difference between their respective transition loci.

The diffuse-rising template for /t/ is slightly wider than it is for /d/ because /t/ bursts tend to be more massive than /d/ bursts. The diffuse-rising template is slightly context-dependent. The template illustrated in Figure 6.11(b) captures alveolar bursts successfully in all contexts except forthe context of the consonant occurring between two back vowels in which case the upper lines of the template must rise more steeply. Both diffuse-rising templates work well for male and female voices.

The final burst rules then are:

BSP = DFL

DSP = DRL(y).$\overline{(DFL.COM(x,y))}$ + 0.5[f,f]DRL(y).$\overline{DFL}$

GSP = COM(x,y)

PSP = DFT

TSP = DRT(y).$\overline{(DFT.COM(x,y))}$ + 0.5[f,f]DRT(y).DFT

KSP = COM(x,y)

where: DFL refers to the diffuse-falling template for lax consonants

DFT refers to the diffuse-falling template for tense consonants

DRL refers to the diffuse-rising template for lax consonants

DRT refers to the diffuse-rising template for tense consonants

COM refers to the compact template

Note: DRL and DRT are written as a function of 'y' to indicate that the particular diffuse-rising template used will depend on the vowel context.

COM is written as a function of 'x' and 'y' to indicate that the compact template used will depend on the sex of the speakers and the vowel context.

6.13.4 - Rules for recognition of plosive consonants in Australian English

Combining the rules for formant transitions and the rules for plosive bursts, the final recognition rules for plosive consonants in Australian English are:

B = BF.BSP + 0.7 BSP + 0.6BF

D = DF.DFP + 0.6DSP + 0.6DF

G = GF.GSP + 0.7GSP +0.6GF

P = PF.PSP + 0.7PSP + 0.6PF

T = TF.TSP + 0.6TSP + 0.6 TF

K = KF.KSP + 0.7KSP + 0.6KF

6.13.5 - Results

Using The rules described in the previous section, it was found that 92% of the plosive consonants tested were correctly and uniquely classified with membership greater than or equal to 0.5. The velar consonants were the most successfully recognised class of sounds with an average of 95% correct recognition. The labial consonants were recognised on average 90% of the time and the alveolar consonants were correctly recognised 91% of the time. In about 6% of all cases consonants were classified correctly but also received a simultaneously high rating in an incorrect category.

6.14 POSSIBLE EXTENSION OF THE PLOSIVE CONSONANT RECOGNITION RULES TO OTHER VOWEL CONTEXTS AND TO OTHER CLASSES OF CONSONANTS

Although the plosive consonant recognition rules were developed from plosives occurring in VCV contexts where the vowels were /ɔ/ and /i/ in various combinations, the rules worked well for plosive consonants surrounded by other front and other back vowels. To accommodate VCV

combinations in which the vowels were central vowels the main modification of the existing rules would be a modification of the rules dealing with formant transitions.

For plosive consonants occurring in XCV or VCX or even XC or CX contexts, where X is a sound (or silence) other than a vowel, the rules given above can be modified to recognise the plosive. For the contexts in which there is a vowel on one side of the plosive both formant transition (between the vowel and the plosive) and burst rules could be used. From the discussions of plosive vowel coarticulation in Section 6.9, it is expected that context dependencies would be mainly reflected in the formant transitions for VCX situations and in the burst in XCV situations. For plosive consonants not adjacent to a vowel (the CX and XC situations) the burst recognition rule would be used because, as was seen in the Section 6.13.3, the three basic types of plosive burst (labial, alveolar and velar) are characterised by three basic burst shapes (diffuse-falling, diffuse-rising and compact).

The spectral template-matching procedure should also be expandable (in modified form) to other classes of consonants. Fricative consonant spectra are also characterised by basic shapes relating to the place of articulation of the consonant (Hughes, Halle and Radley, 1958; Strevens, 1960). And Blumstein and Stevens (1979) have shown that it is possible to classify labial and alveolar nasal spectra correctly 72% of the time using a template-matching procedure.

6.15 CONCLUSION

In this chapter, a set of recognition rules for the plosive consonants has been developed. The success of these rules, which are similar to the recognition rules for Italian plosive consonants described in Chapter 4, indicates that it is possible to recognise the 'same' classes of sounds in different languages by similar methods.

Also investigated were plosive coarticulation patterns in which it was found that extensive coarticulation is seen in the formant transitions between vowels and plosive consonants, but that while coarticulation is strongly marked in velar plosive bursts, it is only slightly noticeable in alveolar bursts and barely noticeable in labial bursts.

The correlates of juncture were also considered. At the phonetic level it was found that the primary effect reflecting juncture was phoneme duration with various other phonetic features acting as secondary effects in this matter.

The most important speaker differences in plosive consonants were found to be the differences between male and female speakers; such differences in the frequency domain followed the same pattern as the male/female second formant differences for vowels. There were other differences between speakers, mainly to do with the degree of coarticulation between a plosive consonant and the adjacent vowels. These differences were not so great however to affect the satisfactory application of the recognition rules.

Chapter 7

THE FOPHO MODEL

7.1 WHITHER (CONTINUOUS) SPEECH RECOGNITION?

In the controversy following the publication of Pierce's letter 'Whither Speech Recognition?' in 1969 (Pierce, 1969,1970; Fleming, 1970; Lea, 1970; Noll, 1970; Samuel, 1970) it was clear that many believed Automatic Speech Recognition to be a task worth pursuing. However, the disillusionment of Pierce's letter, seems to be echoed now, over 10 years later, in the current attitude to the task of continuous speech recognition. At least part of the reason for this is given by Klatt in the concluding remarks of his 'Review of the ARPA Speech Understanding Project' (1978) when he states this '... the nature of the problem is elusive. Even the dimensions of task difficulty have yet to be adequately defined ...'.

As stated previously one of the aims of this thesis has been to investigate some of the parameters which the human perceptual system uses in its very successful decoding of speech, and to incorporate some means of handling these parameters into automatic speech recognition algorithms. The more one investigates the of the human perceptual system's translation continuous speech to a phonetic string, the more complex the problem appears, and the era of machines easily recognizing continuous speech seems further away than ever. In the opinion of the author, it is important that the building of a system to recognize continuous speech not be left until everything is known about the human perception of speech. It means of

course that any system built is hardly likely to be very competent at first but if it is built in such a way that it can be easily modified then recognition will improve as the system is developed. In this way not only is a system being developed but also a framework is provided for testing component recognition algorithms.

In this chapter the design of an acoustic-phonetic decoder operating according to the FOPHO model introduced in Chapter 1 for a Continuous Speech Recognition System is proposed. Possible implementation of such a design is discussed. It is suggested that the design not be implemented as a whole initially but as a series of gradual and integral steps. In such a way the difficulty of the various component steps of the system can be assessed.

From the preceding chapters it has become evident that certain factors have to be considered in the implementation of a system which is to decode continuous speech. From Chapters 3, 4 and 6 consideration of both anticipatory and carryover coarticulatory effects, even across word boundaries, was seen to be crucial to good recognition of consonants. From Chapters 5 and 6 it was seen that normalization across speakers is only necessary in that speakers have to be divided into male and female categories. Adoption of such a procedure implies that no special adaptation of the system to a new structure is necessary, although it can mean that vowels are only recognized, initially at any rate, as being members of a set of two or three vowels. From Chapter 6 it was seen that different speakers make use of the many redundancies in the speech code in different ways and that several conditions of speech parameters can be used for signalling necessary information about a phoneme.

7.2 A FOPHO SYSTEM

Let us consider a foreign phonetician listening to a conversation in Australian English. Let us assume that he has armed himself with a list of the phonemes of Australian English. In order to commence his study he will doubtless interview a native informant in a quiet room, making a tape recording of what is being said. He will ask his informant to name certain objects so that he can get citation from versions of the names of various objects. He might also record conversation between two informants, perhaps asking them to converse on some specific subject. He will then analyse the tape and make a phonetic transcription of it, replaying and relistening to sections of it which he finds difficult to transcribe. From this transcription he can tentatively write down observations on the nature of the phonemes of Australian English noting in particular any features of these phonemes which were peculiar to this dialect. In parallel with this, the phonetician will observe the patterns of prosody (i.e. the intonation, rhythm and stress) patterns of his informant. He will then test to see whether his observations on his informant's speech, apply to other informants from the same dialect group. Observations that apply equally for all informants indicate which features are necessary to the successful production of a sound of that dialect; observations which hold for several informants indicate which features are commonly occurring for speakers of the dialect or indicate legal variations on the production of a given sound. Observations which only hold for the speech of one speaker must be labelled as personal idiosyncracies, and if they do not sound 'odd' to other speakers of the dialect then they must be labelled as legal but not common variations of a sound, or at least the effect must be classed as one which does not interfere with the message content of the speech.

As the phonetician becomes increasingly competent at taking phonetic transcription in his chosen foreign language he will observe more and more modifications and variations to the rules of production for the language. Nevertheless, however diligent he is, it is not necessarily the case that he will discover all the rules pertaining to the production of a given sound of the language i.e. he may discover many rules which are necessary to the production of a sound in a variety of contexts but he may not discover all the rules which are sufficient for the 'proper' production of that sound in all contexts. Thus when he produces the sound himself it will still sound 'foreign'.

Let us consider a design of a Continuous Speech Recognition system which acts in a manner similar to that in which a foreign phonetician acts. This system, as indicated in Chapter 1, will be called FOPHO. The starting point of the system occurs after the system has been equipped with some phoneme recognition rules corresponding to the point after the phonetician has studied intensively some recordings from some informants and is ready to go off and see how well his observations apply for a general cross-section of the population. Like any foreigner the phonetician will have to constantly verify if he has heard what was said correctly.

The aim of a system such as FOPHO is not only to produce a working continuous speech recognition system but also to construct an expandable and learning system i.e. one that can be taught new things and can apply these things to new situations. Such a system is different in concept to a closed system which defines its object-task strictly (such as a spoken chess playing system for example). It is the aim of the FOPHO to learn and be operational at the same time. Its learning should be efficient, making maximal profit from mistake analysis.

In the following sections various features that are seen as necessary for the successful operation of FOPHO are introduced and possible implementation of them described.

7.3 ATTRIBUTES OF FOPHO

Ideally the FOPHO system should have the following attributes if it is to act like a foreign phonetician learning about (the sounds of) a new language.

(1) The system should be very interactive, i.e. it should be able to speak back to the person using it and also clarify points diagrammatically and by written description.

(2) Like a foreigner the system will constantly seek verification to see if it properly understood what the speaker said.

(3) The system must be able to learn from mistakes.

(4) The component programs must be easily modifiable.

(5) It must be easy to add new components to the system.

(6) The quality of component algorithms must be easy to evaluate.

(7) As well as analysing the sounds of the language the system should simultaneously be collecting information about each new speaker.

(8) The system should be able to have higher-level information (i.e. knowledge of syntax, semantics, pragmatics) easily incorporated into it. Or at least it should be able to fit into a general Speech Understanding System.

These ideal attributes will be examined in turn.

7.3.1 A system that answers back

If the FOPHO system is going to even remotely resemble a foreign phonetician it must be equipped with a means of interacting verbally with various users (natives). It is probably readily agreed that the words 'yes' and 'no' are easy to recognize. So the machine* should be taught to recognize (using standard template matching techniques as used in the commercially available 'speaking calculators') each new speaker saying 'yes' and 'no', and perhaps a small number of other words.

The machine would also be equipped with standard interviewing routines comprising a series of questions requiring only 'yes' or 'no' in answer. These routines would:

(a) allow the system to obtain background information about the new speaker;

(b) allow new users of the system to familiarize themselves with the idea of holding a conversation with a machine;

(c) allow the machine to give the person it is 'interviewing' (that is, the native informant in the foreign phonetician analogy) verbal instructions about what to do in the various 'experiments' that it conducts;

(d) conduct experiments.

This concept of set discussions was pioneered by such interactive computer programs as 'Eliza' and 'Doctor' (Weizenbaum, 1968). In these cases however the discussion was not verbal but via teletypes. Such discussions

* The envisaged functioning implementation of the FOPHO system will generally be referred to as 'the machine'. After all, the object of this chapter is to explore a most important aspect of man-machine communication. Perhaps too this is a good place to say that to emphasize that the system is initially being implemented for Australian English, the person that the machine interacts with will be called 'Bruce'.

are not going to fool an expert (or for that matter any layman) into believing that they are speaking to a highly intelligent person but they do give some illusion of intelligence and they do have the overwhelming advantage of being just as usable by the computer novice as by the computer expert. This means that the FOPHO system could conduct its experiments with (or be used by) anyone who was prepared to be reasonably cooperative and just answer 'yes' or 'no' as required or give the machine the recognition tests that it thinks it can handle. (Such tests are described anon).

An introductory conversation between the machine and a new user might then go something like this:

New speaker (after pressing the button marked 'start introductory sequence'): Hello, my name is Bruce

Machine: Spell your name on the keyboard in phonetics

(Bruce does so)

Machine: Your name is Bruce? Bruce do you have an identification number? If so please type it on the keyboard. If not, type 'no number' and you will be assigned a number. Please remember this number.

Will you now please say 'yes' in a normal voice?

Bruce: Yes

Machine: And again

Bruce: Yes

Machine: Will you now please say 'no' in a normal voice?

Bruce: No

Machine: And again

Bruce: No

Machine: I will now ask you some questions. Will you please just answer 'yes' or 'no' as appropriate.

The machine can at this point ask Bruce a series of questions in order to get background information about the speaker. Also it can check its own deductions from measurements made on the speaker's voice as in the following:

Machine: From the pitch of your voice I would say that you are male. Is that correct?

Bruce: Yes

Machine: Are you Australian?

Bruce: Yes

Machine: Do you have a head cold?

Bruce: No

Machine: Do you think that you speak cultivated Australian?

Bruce: No

Machine: Do you think that you speak general Australian?

Bruce: Yes

And so on.

7.3.2 A system that seeks verification

In the above Bruce-Machine conversation the machine is following a set routine of questions. The machine's voice for set conversation pieces could come from pre-recorded speech or, what is possibly better, from synthetic speech. Synthetic speech would definitely be needed when the machine has to produce new words. For example, when the new user has typed his name on the keyboard, the machine says the new speaker's name at the beginning of the next phrase. It would do this by feeding the phonetic transcription to the speech synthesizer which would have rules for the production of a word from a phonetic transcription. Here it is worth mentioning that the typing in of a phonetic transcription presupposes that

either a computer terminal with a phonetic keyboard is available or that input is via a standard keyboard for which some regulated quasi-phonetic input has been devised. A system for doing the latter has recently been developed by Millar, Oasa-Stoycheff, and Wagner (1980).

It should also be noted that for the user to put in words such as his name and also to be able to understand the output from the machine's written transcriptions of what the machine thought it 'heard', the user would need to have some knowledge of phonetic script or some other written pronunciation scheme. This seems to indicate that a totally naive user could not use the system. This is not necessarily so as most people can be shown to write their names phonetically easily, and such users need not concern themselves with written phonetic output. Users who wish to interact extensively with the system would find that it was worth the trouble to acquire a working knowledge of broad phonetic script. Or, but this seems to be complicating matters unnecessarily, a text-to-speech system (Lesmo, Mezzalama and Torasso, 1978) could be incorporated into the FOPHO system. Such a refinement might be hardly necessary in the developmental stages of FOPHO but it could be profitably incorporated in a final system in order to make the system available to a large number of users.

The major use of synthetic speech in the system, however, would be to allow the system to reproduce a sound, word, or phrase that it thinks it has 'heard'. Like all foreigners the system will constantly seek verification from the person with whom it is speaking, checking to see if it correctly understood what the speaker means to say. This leads to the following (rather tedious) type of conversation:

Machine: Say a word, please Bruce

Bruce: Dog

Machine: Did you say /bɔk/?

Bruce: No

Machine: Did you say /bʊk/, then?

Bruce: No

Machine: However is the vowel in the word /bʊk/ the same as the vowel in the word that you said?

Bruce: Yes

Machine: Perhaps you said /bʊg/? Is the last sound in the word correct?

Bruce: Yes

.

.

.

The machine can thus go on proposing its second best (and substantially its further best guesses) until it either gets the word right or, after a certain number of guesses, it will write (in phonetics) all its 'remote' guesses on a teletype screen and ask the user to indicate by some means which, if any, of these guesses is correct.

7.3.3 A system that learns from its mistakes

Any system such as FOPHO should be able, particularly at the developmental and learning stages, to give the user quick and accurate information about its method of arriving at a conclusion so that the user can discover where it is making mistakes or running into difficulties.

For a start, let us consider how the dialogue of the last section might continue. Suppose the machine is allowed to propose all its 'guesses' which have a fuzzy rating above a certain limit, say 0.7.

Further suppose its fuzzy rating for /d/ as the first sound of the word is 0.4. Thus it can never guess the word 'dog' /dɒg/ with a fuzzy rating greater than 0.4. So after its proposals of /bɔk/, /bɒk/ and /bɒg/ all of which, say, had fuzzy ratings greater than or equal to 0.7 it might continue its conversation as follows:

Machine: Please write the word on the keyboard in phonetics

(Bruce does so)

Machine: My algorithm for /d/ could be incorrect

To make this deduction the system simply finds that there is a mismatch between its 'guesses' and the sound the user has told it that he said. Actually the system could be equipped with the capacity to make various verbal deductions. For example it could continue:

Machine: I was able to deduce that the first sound was a plosive consonant, but I could not get the place of articulation correct.

In analysing its mistakes the system should be equipped with the capacity to retrace its steps in arriving at a conclusion with the object of discovering at what level the mistake was made. This is possibly best done by comparing the entries in a table of features of the sound that was guessed by the machine with the entries in a table of features of the sound that the user claims (when he types the word in at a terminal using phonetic script) he said.

Of course the system could still make other mistakes, for example in the segmentation of the word into phonemes. If it found a mismatch between the number of phonemes it found and the number of phonemes it should have found it could say: 'I made an error in segmentation'.

At the same time as the machine is saying all this it should display graphically approximately the same information that is being transmitted verbally. Although it is easiest to assimilate spoken information (Ochsman and Chapnais, 1974) it is useful to have a diagrammatic display particularly if a hard-copy version of this can also be obtained. For example, a flowchart type illustration of the method of deducing the identity of the word spoken could be displayed. At the bottom of the screen instructions could be given for obtaining more detailed information about different parts of the method. Thus on requesting parameter information the machine would print the actual measured values of parameters used in arriving at its various decisions.

7.3.4 An easily modifiable system

A useful adjunct to having the system carry out a mistake analysis would be the facility to modify simply the programs by which the recognition algorithms are implemented. It would be particularly useful if such tuning of the system could be carried out interactively. In general two types of modification could be needed:

(1) Modification of the rule structure of a recognition algorithm including the introduction of new recognition rules.

(2) Modification of the fuzzy vector of breakpoints of a parameter of the recognition algorithms.

In Chapter 2 there is a description of a finite fuzzy automaton. From that description it can be seen that modification of the rule structure is a matter of passing new input parameters to the subroutine (AUTMN). To do this the user could select the option to modify the rules. The machine would then interrogate the user by displaying questions on the screen which

would ask the user to input new values at the keyboard. To jog the memory of the user the present rule structure might be displayed on the screen (we assume that a plotting screen is available).

Similarly modification to fuzzy boundaries could be done by the user selecting the appropriate modification option and the machine responding with a series of questions, the user-supplied responses to which will allow the system to carry out the appropriate changes to the appropriate subroutine.

7.3.5 An easily extendable system

So far fairly simple word recognition procedures for the FOPHO system have been discussed. As it is basically intended to be a system which can (eventually) decode continuous speech, it is most important that the system be reasonably adept at handling phrases as well as words. Phrase recognition involves all the difficulties of word recognition plus several other problems such as word boundary detection and vowel reduction. As well as this there is a greater chance of errors being made at the segmentation stage. It is likely that not all sounds will be detected and also that additional phonemes to those that are present will also be hypothesised.

Basically phrase recognition procedures would be similar to the word recognition procedures. Again the system would attempt to segment the incoming speech waveform into phonemes of various general classes (initially sonorant/nonsonorant). Then each sound would be further classified first into classes indicating manner of production and then according to place of production. However, when recognizing phrases the system would also have to be alert for indications of word boundaries.

Long periods of silence are very definite indicators of a word boundary but as was seen in Chapter 6, word boundaries are often not accompanied by silence and the only indicator at the acoustic-phonetic level of a word boundary would be the possible presence of word-initial and/or word final forms of various phonemes. When the system encountered such a phoneme it would assign a word boundary with a fuzzy weighting of its plausibility.

All the procedures already outlined for testing and correcting the machine's capacity for word recognition could be used for phrase recognition. The user could say a phrase which the system would attempt to decode and would then 'say' (in synthetic speech) what it 'heard' and also test other likely variants on the screen. Thus:

Machine: Say a sentence

Bruce: I eat chips, do you?

Machine: Did you say /ait tʃɪps deju/. I have listed on the screen my guesses of what you said. Could you underline the sounds that are correct with the light pen. Could you indicate with vertical lines where the word boundaries should occur. Have you finished?

Bruce: Yes

Machine: I will now ask you questions about the sounds I did not recognise.

.

.

.

Again a rather tedious conversation but again the person using the machine would be able to find out the mistakes that the machine is making and these could be corrected using the interactive modification routines.

It should be noted here that it does not matter greatly if the system is not detecting all the word boundaries as the position of a word boundary is often not indicated at the phonetic level (see Chapter 6) of speech understanding. Semantic, syntactic and prosodic processing will generally define the word boundaries more accurately. This brings us to the question of how such processing can be incorporated into the FOPHO system. Each type of processing would be done in a modular fashion with a blackboard arrangement similar to that used in the HEARSAY II speech understanding system (Erman et al, 1980) for controlling the methods of operation of the various processes. For example, the blackboard controller would send incoming waveforms simultaneously to the acoustic phonetic and prosodic processors and as these processors return output the controller would send this output to other processors. Further discussion of the interacting between various processors will be given in a later section.

7.3.6 A system in which the components can easily be tested

It will be becoming increasingly evident that the FOPHO system would, in its early stages of development and particularly before higher-level (i.e. semantic and syntactic) knowledge has been incorporated into it, be basically a system which would allow those interested in automatic speech recognition to evaluate interactively the quality of their algorithms for acoustic-phonetic processing. The design of such algorithms is, as is illustrated by the four preceding chapters, an involved matter of examining the literature on the type of classification being attempted to discover the parameters which are most likely to be relevant, followed by the examination of a large amount of data to find if such parameters can be measured easily (and automatically) and if they give sufficient information to allow unique classification, and if they do not, to select other

parameters which will achieve the classification. The time taken on this process of algorithm design, particularly when one is aiming at designing algorithms which are to be speaker-independent and sufficiently robust to classify a sound occurring in every possible sound context is considerable. Rather than waiting until all components of the acoustic-phonetic processor are designed it is probably wiser to implement a system which initially only recognizes a subset of all the sounds in a language. Indeed Koster (1980) has designed a speech recognition system which only recognizes speech in an artificial language, the sounds of which are a subset of the phonemes of German.

There are advantages to implementing acoustic-phonetic recognition in a piecemeal fashion. It implies that the system must be modular in design with a facility for easy incorporation (or deletion) of recognition algorithms. Also the system's operation on a small number of recognition algorithms can be studied systematically. For example, one could construct a system which contained only algorithms which would theoretically allow one to classify sounds as either vowels, plosives or 'something else'. Thus if the algorithms are working correctly conversations like this could take place:

Machine: Say a sentence

Bruce: A cat bit a dog

Machine: You said /akæt bɪt a dɒg/

Bruce: Yes

Machine: Say another sentence

Bruce: A cat hit a frog

Machine: I cannot classify all the sounds you said. I will put a buzz sound which I write as an asterisk on the screen instead of the sounds I don't know. So I think you said /akæt *ɪt a**ɒg/.

Such an implementation of limited subset of recognition algorithms allows the user to focus on the problem of the system's capacity for classifying a sound in one category with a high fuzzy rating while simultaneously classifying that sound in other categories with a low fuzzy rating. In the plosive-vowel-anything else system one would not want a consonant to be rated both as 'a plosive' and 'anything else', for example. Also in this plosive-vowel-anything else system one could evaluate the word boundary algorithms which were developed in Chapter 6.

7.3.7 A system with on-going speaker adaptation

In Section 7.3.1 it was suggested that when a new speaker is introduced to the system, the system would allot that speaker an identification number by which that speaker could identify himself to the machine on future occasions. The machine would use this identification number to access file containing information about the speaker. This file would contain information that the system got by asking the user questions requiring yes/no answers and also information that it has gathered from an analysis of its mistakes.

In Chapter 5 the need for gathering on-line information about a speaker's vowels was discussed in detail. Once the FOPHO system has 'heard' the speaker produce all the vowels of the language and 'realized' what each vowel was meant to be it could estimate the (non-overlapping) regions corresponding to this speaker's vowels in formant space. The file for each speaker would consist of an initially blank array the various sections of which are to be filled with parameter values from measurements on sounds whose identity is known. Thus in the conversation between Bruce and the machine in Section 7.3.2, when the machine finds that when finally

it has correctly identified the vowel in the word 'dog' it would put the values of the first form formants of the vowel into the section of the user file reserved for 'formant values of the vowel /ɒ/'.

After a sound has been classified as being in the category 'vowel' the system controller would activate the algorithm which decides to which subset of the vowels the unknown vowel must belong and pass this information along with some parameter information back to the controller. The controller would then initiate a search of its user file to see if it has sufficient information to decide definitely which vowel the unknown vowel would most probably be on the basis of past productions of the hypothesized vowels. When it has been found by user verification, what the identity of the vowel actually was the parameter values of that vowel are also added to the user file in the appropriate positions. And so on until sufficient information is gathered about each vowel that the system makes sufficiently few mistakes in vowel recognition for that speaker to be confident that he is very familiar with the speaker's vowels.

On-line collection and storage of information for each speaker need not be restricted to vowels. In Chapter 6 it was seen that speakers can produce the plosive consonants in various different ways. In the user file one might store, for example, information about the type of coarticulation effects displayed in various plosive sounds, whether or not aspiration usually accompanies the production of a particular plosive in a particular context, etc. As well as storing vowel and consonant parameter information the system might store information about the speaker's overall speaking rate, the type of intonation patterns he uses and other such prosodic phenomena.

7.3.8 A speech recognition system that can grow into a speech understanding system

The FOPHO system could be extended from being merely a system which recognizes sound combinations of a language to one which understands speech in that language. Again it is useful to have the system approach the problem in a way similar to the way in which a foreigner would learn to understand a foreign language.

Let us consider how the Marslen-Wilson and Welsh model of word recognition (1978) which was discussed in Chapter 1 could be implemented as part of the FOPHO system. According to that model the incoming speech is continually being coded phonetically, this phonetic code is recoded phonologically, and then a search of the lexicon is made to find what words of the lexicon are activated by the initial sounds of an unknown word. The remaining phonemes of a word are then inspected by a verification procedure to see which of the activated words the unknown word is or, if no match is found, the unknown word is classed as being an 'unfamiliar' word.

To actually implement this system for the recognition of continuous speech would be a massive task. However, a foreigner will initially only have a small vocabulary and will at best only pick up a few words. A FOPHO system could have a small basic lexicon and phonetic representatives of each word in the lexicon would be stored.

As a sentence is spoken the incoming speech would be segmented and each phoneme would be classified. As soon as a sound has been classified as well as possible, the result would be sent to the controller which would then search the lexicon for words beginning with that sound and perhaps for words beginning with the 'second-best' guess at that sound. The system

would then initiate verification procedures to attempt to find out which of the possibilities the incoming word is. These verification routines could be derived from the various recognition algorithms. Each activated word would be investigated. As further results both from the recognition and the verification procedures became available the search would be concentrated on a decreasingly small number of possibilities.

Of course the matching up procedures between the phonetic version of each word stored in the lexicon and the incoming speech would have to take account of effects such as vowel reduction, ellision, and apocope. The problem of word boundary detection could be approached in several ways. As has been noted, some indicators of a word boundary come from the acoustic-phonetic and prosodic decoders. A knowledge of non-allowable within-word sound combinations (e.g. /vp/) could be built into the system. If such a combination was encountered it would mean that there was a word boundary between them. Finally if a combination of incoming sounds has been found to be a word of the lexicon, it can be assumed that the following sounds are either a suffix or the start of a new word.

The system could also be equipped with syntactic knowledge which would allow it it 'understand' something of the grammatical structure and the message of the incoming speech. Semantic knowledge might be incorporated by having each word of the lexicon cross-referenced to other words with which it is commonly associated.

Although the task of incorporating higher level knowledge into the system is large, it can be done in slow stages. Each stage can be tested by interactive procedures and the system can be trained to analyse its mistakes at each stage. Even though it might only recognize very few words out of each sentence, it will be doing as well as foreigners do when

learning a new language. And like the foreigner the FOPHO system could learn from its mistakes.

7.4 CONCLUSION

A system for continuous speech recognition has been proposed and the desirable attributes of the system have been discussed in detail. This was done to demonstrate that a continuous speech recognition system can be feasibly envisaged. By taking the learning processes of a foreign phonetician as a model for the system, it is tacitly acknowledged that the system would not be meant to achieve perfect sound recognition (or word understanding). Rather a system which achieves some (where 'some' would initially mean 'a very little' and probably, after extensive system development, 'not inconsiderable') recognition of what was said would be the goal. The great benefits of such a design arise from the fact that it offers a means of tackling the continuous speech recognition problem with a system which although it is not expected to perform well at first should be able to evolve easily into a better-performing system as better recognition algorithms are added to it. Also the system is aimed at being easy to use. This will mean that system tests can be carried out by a large number of people. As more people use the system, various ways in which the design ought to be modified would become clearer.

In parallel with the building of a system such as the FOPHO system the design of new and comprehensive phoneme recognition algorithms would have to be undertaken.

Chapter 8

CONCLUSION

8.1 ACHIEVEMENTS IN RELATION TO GOALS

On page 2 the goals of the work done in this thesis were stated. The first goal was to consider the general problem of phoneme recognition with a view to discovering why automatic phoneme recognition systems did not obtain very accurate recognition. In Chapter 1 early stages of speech recognition (up to the level of word recognition) were surveyed in detail. It was seen that automatic recognition schemes (described in Section 1.5) only accounted for a limited number of the speech production phenomena which the human perceptual mechanism is aware of and uses in its speech decoding process. Thus it is not surprising that the automatic speech recogniton algorithms did not achieve high recognition rates. However, the amount of work involved in the development of the algorithms described in Chapters 3-6 makes it clear that automatic recognition algorithms which account for a large variety of speech phenomena are time-consuming to develop and even then do not achieve perfect recognition. This is the reason, then, that the FOPHO model of automatic speech recognition is proposed and discussed in detail. The FOPHO system should be a system which is continuously learning and adapting (in a manner similar to that reasonably well understood human method of learning - human language learning) and thereby decreasing the amount of work that the designer of an automatic acoustic-phonetic decoder has to do.

The second goal was to find new ways to incorporate, into automatic acoustic-phonetic recognition algorithms, information about the speech code that has been discovered through research in a variety of disciplines. This was done with some success. For example, the inclusion of coarticulation effects led to greatly improved consonant recognition scores. Nevertheless there are many effects which are as yet imperfectly understood in the disciplines in which they are usually studied, and thus it is difficult to incorporate results about them into automatic recognition algorithms. An example of this is the lack of any comprehensive psycholinguistic theory of information integration as regards integrating information about various features of a sound in such a way as to achieve phonetic recognition of that sound. Thus information integration in automatic acoustic-phonetic recognition systems is done in an *ad hoc* fashion.

The third goal which was to clarify the quality of acoustic-phonetic recognition needed for interaction with other components of a speech recognition system was achieved fully only within the context of the FOPHO model. A FOPHO model can tolerate reasonable mistakes in recognition if adequate facilities (both verbal and written) with the speaker.

The fourth goal was to investigate several specific problems for acoustic-phonetic recognition of continuous speech. While the problems of juncture and coarticulatory word boundary phenomena were investigated only for vowel-plosive consonant-vowel contexts, the results are such that they should be easily able to be generalized to other phonetic contexts. The problem of speaker normalization was also considered and it was found that although account must be taken of the sex of the speaker acoustic-phonetic recognition algorithms could be designed to be otherwise speaker independent.

8.2 CONCLUSION

In summary, this thesis has provided several phoneme recognition algorithms which are capable of achieving good recognition results in cases of continuous conversational speech without any necessary prior knowledge about the speaker's voice. The problems of coarticulation, speaker normalization, and word boundary phenomena were also investigated. As well, a model (FOPHO) of an intelligent acoustic-phonetic decoder has been proposed.

REFERENCES

Abramson, A S and Lisker, L, 'Voice onset time in stop consonants: Acoustic analysis and synthesis', Proc 5th Int Congress of Acoustics, Liege, 1965.

Ali, L, Gallagher, T, Goldstein, J and Daniloff, R, 'Perception of coarticulated nasality', J Acoust Soc Amer, 49, pp 538-540, 1971.

Alinat, P, 'Edute du trait permettant de distinguer entre les 3 classes de consonnes explosives PB, TD, KG', Textes des exposes de 9emes Journees d'Etude sur la Parole, Lanion, France, pp 297-303, May-June, 1978.

Atal, B S and Hanauer, S L, 'Speech analysis and synthesis by linear prediction of the speech wave', J Acoust Soc Amer, 50, pp 637-655, 1971.

Becker, R W and Poza, F, 'Acoustic phonetic research in speech understanding', IEEE Trans Acoust Speech Sig Proc, Vol ASSP-23, pp 416-426, 1975.

Bellman, R E and Giertz, M, 'On the analytic formalism of the theory of fuzzy sets', Infor Sci, 5, pp 149-156, 1973.

Bengeurel, A P and Cowan, H A, 'Coarticulation of upper lip protrusion in French', Phonetica, 30, pp 41-45, 1974.

Bernard, J R L, 'Some measurements of some sounds of Australian English', PhD thesis, University of Sydney, Sydney, 1967a.

Bernard, J L R, 'Length and the identification of Australian English vowels', AUMLA, 27, pp 37-58, 1967b.

Bernard, J R L, 'On the uniformity of Australian English', Orbis, Vol

XVIII, No 1, 1969.

Bernard, J R L, 'Toward the acoustic specification of Australian English', Zeitschrift fur Phonetik, Band 23, Heft 2/3, pp 113-128, 1970.

Blumstein, S E, Stevens, K N and Nigro, G N, 'Property detectors for bursts and transitions in speech perception', J Acoust Soc Amer, 61, 5, 1977.

Blumstein, S E and Stevens, K N, 'Acoustic invariance in speech production: Evidence from measurements of the spectral properties of stop consonants', J Acoust Soc Amer, 66, pp 1001-1017, 1979.

Blumstein, S E and Stevens, K N, 'Perceptual invariance and onset spectra for stop consonants in different vowel environments', J Acoust Soc Amer, 67, pp 648-662, 1980.

Burgess, O N, 'A spectrographic investigation of some Australian vowel sounds', Language and Speech, 11, pp 129-137, 1968.

Butcher, A and Weiher, E, 'An electropalatographic investigation of coarticulation in VCV sequences', J Phonet, 4, pp 59-74, 1976.

Chomsky, N and Halle, M, 'The sound pattern of English', New York:Harper and Row, 1968.

Christie Jr, W M, 'Some cues for syllable juncture perception in English', J Acoust Soc Amer, 55, pp 819-821, 1974.

Christie Jr, W M, 'Some multiple cues for juncture in English', General Linguistics, 17, pp 213-222, 1977.

Cole, R A, 'Listening for mispronunciation: A measure of what we hear during speech', Perception and Psychophysics, 13, pp 153-156, 1973.

Cole, R A, Jakimik, J and Cooper, W E, 'Segmenting speech into words', J Acoust Soc Amer, 67, pp 1323-1332, 1980.

Cole, R A and Scott, B, 'The phantom in the phoneme: Invariant cues for stop consonants', Perception Psychophysics, 15, pp 101-107, 1974.

Cole, R A and Scott, B, 'Towards a theory of speech perception', Psychol Rev, 81, pp 348-374, 1974.

Cooley, J W and Tukey, J W, 'An algorithm for the machine calculation of complex Fourier series', Math Comp, 19, pp 297-301, 1965.

Cooper, F S, Delattre, P C, Liberman, A M, Borst, J M and Gerstman, L J, 'Some experiments in the perception of synthetic speech sounds', J Acoust Soc Amer, 24, pp 597-606, 1952.

Cutting, J and Rosner, B S, 'Categories and boundaries in speech and music', Perception and Psychophysics, 16, pp 564-570, 1974.

Daniloff, R G and Hammarberg, R E, 'On defining coarticulation', J Phonet, 1, pp 239-248, 1973.

Datta, A K, Ganguli, N R, Ray, S and Mukherjee, B, 'Computer recognition of plosive speech sounds', IEEE Conference on Computers, Session on Pattern Recognition and Learning Methods, pp 122-134, February 1978.

Davis, H K, Biddulph, R and Balashek, S, 'Automatic recognition of spoken digits', J Acoust Soc Amer, 24, pp 637-642, 1952.

Delattre, P C, Liberman, A M and Cooper, F S, 'Acoustic loci and transitional cues for consonants', J Acoust Soc Amer, 27, pp 769-773, 1955.

Demichaelis, P, De Mori, R, Laface, P and O'Kane, M, 'Computer recognition of stop consonants', Proceedings IEEE Conference on Acoustics, Speech and Signal Processing, Washington, 1979.

De Mori, R, 'Syntactic recognition of speech patterns', in 'Syntactic Pattern Recognition, Applications', K S Fu (ed), New York: Springer-Verlag, 1977.

De Mori, R, Private communication, 1980.

De Mori, R and Laface, P, 'Use of fuzzy algorithms for phonetic and phonemic labelling of continuous speech', IEEE Trans on Pattern Analysis and Machine Intelligence, Vol PAMI-2, No 2, pp 136-148, 1980.

De Mori, R, Laface, P and Piccolo, E. 'Automatic detection and description of syllabic features in continuous speech', IEEE Trans Acoust Speech Sig Proc, Vol ASSP-24, pp 365-378, 1976.

De Mori, R, Laface, P and Torasso, P, 'Automatic recognition of liquids and nasals in continuous speech', Proc 1977 IEEE Conference on Acoustics, Speech and Signal Processing, Hartford, Conn, pp 644-647, 1977.

De Mori, R, Rivoira, S and Serra, A, 'A speech understanding system with learning capability', Proc 4th Internat Joint Conf on Artificial Intelligence, Tibilsi, USSR, September 1975.

Denes, P B and Matthews, M V, 'Spoken digit recognition using time-frequency pattern matching', J Acoust Soc Amer, 32, pp 1450-1455, 1960.

Dorman, M, Studdert-Kennedy, M and Raphael, L, 'Stop consonant recognition: Release bursts and formant transitions as functionally equivalent,

context sensitive cues', Perception and Psychophysics, 22, pp 109-122, 1977.

Eimas, P D and Corbit, J D, 'Selective adaptation of linguistic feature detectors', Cognitive Psychology, 4, pp 99-109, 1973.

Eimas, P D and Miller, J L, 'Effects of selective adaptation on the perception of speech and visual patterns: Evidence for feature detectors', in 'Perception and Experience', R D Walk and H L Pick (eds), New York:Plenum, 1978.

Erman, L D, Hayes-Roth, F, Lesser, V R and Reddy, D R, 'The Hearsay-II speech-understanding system: Integrating knowledge to resolve uncertainty', Computing Surveys, 12, pp 219-253, 1980.

Fant, G, 'A note on vocal tract size factors and non-uniform F-pattern scalings', Speech Transmission Lab Quarterly Progress and Status Report, 4, pp 22-30, 1966.

Fant, G, 'The nature of distinctive features', in 'To honor Roman Jacobson: Essays on the occasion of his seventieth birthday', The Hague:Mouton, 1967

Flanagan, J L, 'Speech Analysis Synthesis and Perception', New York:Springer-Verlag, 1972.

Fleming, L, 'Implications of speech recognition studies', J Acoust Soc Amer, 47, p 1612, 1970.

Forgie, J W and Forgie, C D, 'Results obtained from a vowel recognition computer program', J Acoust Soc Amer, 31, pp 1480-1489, 1959.

Forster, K I, 'Accessing the mental lexicon', in 'New approaches to

language mechanisms', R J Wales and E C T Walker (eds), Amsterdam:North Holland, 1976.

Foss, D J and Blank, M A, 'Identifying the speech codes', Cognitive Psychology, 12, pp 1-31, 1980.

Foss, D J and Dowell, B E, 'High speed memory retrieval with auditory presented stimuli', Perception and Psychophysics, 9, pp 465-468, 1971.

Foulkes, J D, 'Computer identification of vowel types', J Acoust Soc Amer, 33, pp 7-11, 1961.

Fowler, C A, 'Coarticulation and theories of extrinsic timing', J. Phonet, 8, pp 113-133, 1980.

Fry, D B, Abramson, A S, Eimas, P D and Liberman, A M, 'The identification and discrimination of synthetic vowels', Language and Speech, 5, pp 171-189, 1962.

Fujimura, O, 'Analysis of nasal consonants', J Acoust Soc Amer, 34, pp 1865-1875, 1962.

Fujisaki, H, Tanaka, H and Higuchi, N, 'Analysis and feature extraction of voiced stop consonants in Japanese', Trans Committee on Speech Research, Acoustical Society of Japan, No S79-12, pp 89-96, May 1979.

Gaines, B R, 'Foundations of fuzzy reasoning', Int J Man-Machine Studies, 8, pp 623-668, 1976.

Gaines, B R and Kohout, J L, 'The fuzzy decade: A bibliography of fuzzy systems and closely related topics', Int J Man-Machine Studies, 9, pp 1-68, 1977.

Gay, T, 'A cinefluorographic study of vowel production', J Phonet, 23, pp 255-266, 1972.

Gay, T, 'Articulatory movements in VCV sequences', Haskins Lab Status Report, SR-49, pp 121-147, 1977.

Gerstman, L J, 'Classification of self-normalised vowels', IEEE Trans on Audio and Electroacoust, AU-16, pp 78-80, 1968.

Goldberg, H G, 'Segmentation and labelling of speech: A comparative performance evaluation', PhD thesis, Department of Computer Science, Carnegie-Mellon University, Pittsburgh, 1975.

Halle, M, Hughes, G W and Radley, J P A, 'Acoustic properties of stop consonants', J Acoust Soc Amer, 29, 1, pp 107-116, 1957.

Hammarberg, R, 'The metaphysics of coarticulation', J Phonet, 4, pp 353-363, 1976.

Hanley, T D and Andrews, M L, 'Some acoustic differences betweed educated Australian and General American dialects', Phonetica, 17, pp 241-249, 1967.

Harris, K S, Hoffman, H S, Liberman, A M, Delattre, P C and Cooper, F S, 'Effect of third-formant transitions on the perception of voiced stop consonants', J Acoust Soc Amer, 30, pp 122-126, 1958.

Hoffman, H S, 'Study of some cues in the perception of voiced stop consonants', J Acoust Soc Amer, 30, pp 1035-1041, 1958.

Itakura, F, 'Minimum prediction residual principle applied to speech recognition', IEEE Trans Acoust Speech Sig Proc, ASSP-23, pp 67-72, 1975.

Jelinek, F, 'Continuous speech recognition by statistical methods', Proc IEEE, 64, pp 532-556, 1976.

Kent, R D and Minifie, F D, 'Coarticulation in recent speech production models', J Phonet, 5, pp 115-133, 1977.

Klatt, D H, 'Vowel lengthening is syntactically determined in a connected discourse', J Phonet, 3, pp 129-140, 1975.

Klatt, D H, 'Linguistic uses of segmental duration in English: Acoustic and perceptual evidence', J Acoust Soc Amer, 59, pp 1208-1221, 1976.

Klatt, D H, 'Review of the ARPA Speech Understanding Project', J Acoust Soc Amer, 62, pp 1345-1366, 1977.

Klatt, D H, 'Speech perception: A model of acoustic-phonetic analysis and lexical access', J Phonet, 7, pp 279-312, 1979.

Klatt, D H and Stevens, K N, 'On the automatic recognition of continuous speech: Implications from a spectrogram-ready experiment', IEEE Trans Audio Electroacoustic, AV-21, pp 210-217, 1973.

Kohonen, T, 'Computer Addressable Memories', New York:Springer-Verlag, 1980.

Koster, J-P, 'Verstandlichkeit geflusterter kunstlicher Worter', presented at the 10th International Congress on Acoustics, Sydney, 1980.

Kuehn, D P and Moll, K L, 'Perceptual effects on forward coarticulation', J Speech and Hearing Research, 15, pp 654-664, 1972.

Kuehn, D P and Moll, K L, 'A cineradiographic study of VC and CV articulatory velocities', J Phonet, 6, pp 303-320, 1976.

Kuhn, G M, 'On the front cavity resonance and its possible role in speech perception', J Acoust Soc Amer, 58, pp 428-433, 1975.

Kuhn, G M, 'Stop consonant place perception with single-formant stimuli: Evidence for the role of the front-cavity resonance', Haskins Lab Status Report, SR-57, pp 113-114, 1979.

Ladefoged, P and Broadbent, D E, 'Information conveyed by vowels', J Acoust Soc Amer, 29, pp 8-104, 1957,

Laface, P, 'A formant tracking system toward automatic recognition of speech', Signal Processing, 2, pp 113-129, 1980.

Larkey, L S, Wald, J and Strange, W, 'Perception of synthetic nasal consonants in initial and final syllable position', Perception and Psychophysics, 23, pp 299-312, 1978.

LaRiviere, C, Wintz, H and Herriman, E, 'Vocalic transitions in the perception of voiceless initial stops', J Acoust Soc Amer, 57, pp 470-475, 1975.

Lasky, R E, Syrdal-Lasky, A and Klein, R E, 'VOT discrimination by four to six and a half month old infants from Spanish environments', J Exp Child Psychology, 20, pp 215-225, 1975.

Lea, W A, Medress, M F, and Skinner, T E, 'A prosodically guided speech understanding system', IEEE Trans Acoust Speech and Signal Processing, ASSP-23, pp 30-38, 1975.

Lehiste, I, 'An acoustic-phonetic study of internal open juncture', Suppl ad Phonetica, 5, pp 1-54, 1960.

Lehiste, I and Shockey, L, 'On the perception of coarticulation effects in English VCV syllables', J Speech Hearing Res, 15, pp 500-506, 1972.

Lesmo, L, Mezzalama, M and Torasso, P, 'A text-to-speech translation system for Italian', Int J Man-Machine Studies, 10, pp 569-591, 1978.

Liberman, A M, 'The grammars of speech and language', Cognitive Psychology, 1, pp 301-323, 1970.

Liberman, A M, Cooper, F S, Shankweiler, D P and Studdert-Kennedy, M, 'Perception of the speech code', Psychological Review, 74, pp 431-459, 1967.

Liberman, A M, Delattre, P C and Cooper, F S, 'The role of selected variables in the perception of the unvoiced stop consonants', Amer J Psychol, 65, pp 497-516, 1952.

Liberman, A M, Harris, K S, Hoffman, H S and Griffith, B C, 'The discrimination of speech sounds within and across phoneme boundaries', J Exp Psych, 54, pp 358-368, 1957.

Liberman, A M, Harris, K S, Kinney, J A and Lane H, 'The discrimination of relative onset time of the components of certain speech and non-speech patterns', J Exp Psych, 61, pp 379-388, 1961.

Lindblom, B and Studdert-Kennedy, M, 'On the role of formant transitions in vowel recognition', J Acoust Soc Amer, 42, pp 830-843, 1967.

Lisker, L and Abramson, A S, 'Some effects of context on voice onset time in English stops', Language and Speech, 19, p 1-28, 1967.

Lowerre, B T, 'The HARPY speech recognition system', PhD thesis, Department of Computer Science, Carnegie-Mellon University, Pittsburgh, 1976.

McClean, M, 'Forward coarticulation of velar movement and marked junctural boundaries', J Speech and Hearing Research, 16, pp 286-296, 1973.

Malmberg, B, 'The phonetic basis for syllable division', Studia Linguistica, 9, pp 80-87, 1955.

Marslen-Wilson, W D, 'Sentence perception as an interactive parallel process', Science, 189, 226-228, 1975.

Marslen-Wilson, W D and Welsh, A, 'Processing interactions and lexical access during word recognition in continuous speech', Cognitive Psychology, 10, pp 29-63, 1978.

Massaro, D W, 'Perceptual units in speech recognition', J Exp Psych, 102, pp 199-208, 1974.

Massaro, D W and Cohen, M M, 'The contribution of fundamental frequency and voice onset time to the /zi/-/si/ distinction', J Acoust Soc of Amer, 60, pp 704-19, 1976.

Medress, M F, Cooper, F S, Forgie, J W, Green, C C, Klatt, D H, O'Malley, M H, Newburg, E P, Reddy, D R, Ritia, B, Shoup-Himmel, J E, Walker, D E and Woods, W A, 'Speech understanding systems: Report of a Steering Committee', Sigart Newsletter 62, pp 4-7, 1977.

Massaro, D W and Oden, G C, 'Evaluation and integration of acoustic features in speech perception', J Acoust Soc Amer, 67, 3, pp 996-1013, 1980.

Mermelstein, P, 'On detecting nasals in continuous speech', J Acoust Soc Amer, 61, pp 581-587, 1977.

Mermelstein, P, 'Difference limens for formant frequencies of steady-state and consonant-bound vowels', J Acoust Soc Amer, 63, pp 572-580, 1978.

Millar, J B, 'An audio-visual waveform editing progam', Proc DECUS, Canberra, pp 1335-1336, 1978.

Millar, J B, Oasa-Stoycheff, H, and Wagner, M,, 'Towards modelling of speaker characteristics', J Acoust Soc Amer, 67, Suppl No 1, p S94, 1980.

Miller, G A and Nicely, P E, 'An analysis of perceptual confusions among some English stop consonants', J Acoust Soc Amer, 27, pp 338-352, 1954.

Miller, J L and Eimas, P D, 'Studies on the selective tuning of feature detectors for speech', J Phonet, 4, pp 119-127, 1976.

Mitchell, A G and Delbridge, A, 'The Speech of Australian Adolescents: A Survey', Angus and Robertson, Sydney, 1965.

Moll, K L and Daniloff, R G, 'Investigation of the timing of velar movements during speech', J Acoust Soc Amer, 50, pp 678-684, 1971.

Morais, J, Cary, L, Alegria, J and Bertelson, P, 'Does awareness of speech as a sequence of phones arise spontaneously', Cognition, 7, pp 323-331, 1979.

Morton, J and Long, J, 'Effect of word transitional probability on phoneme identification', J of Verbal Learning and Verbal Behaviour, 15, pp 43-51, 1976.

Noll, A M, 'Whither speech production?', J Acoust Soc Amer, 47, pp 1614-1616, 1970.

Oasa, H, 'Regional Variations in Australian English', M.A. thesis, Aust Nat Univ, Canberra, 1980.

Ochsman, R B and Chapanis, A, 'The effects of 10 communication modes on the behaviour of teams during co-operative problem-solving, Int J Man-Machine Studies, 6, pp 579-619, 1974.

Oden, G C, 'Integration of place and voicing information in the identification of synthetic stop consonants', J Phonet, 6, pp 83-93, 1978.

Ohman, S E G, 'Coarticulation in VCV utterances: Spectrographic measurements', J Acoust Soc Amer, 39, pp 151-168, 1966.

O'Kane, M, 'Linear prediction in speech analysis', presented at 2nd Australian Institute of Physics Congress, Sydney, 1976.

O'Kane, M, 'The use of prosody in speech recognition', presented at First Australian Language and Speech Conference, Melbourne, 1977.

O'Kane, M, 'New approaches to the acoustic-phonetic component of a speech recognition system', Australian Computer Science Communications, 2, pp 69-83, 1980.

O'Malley, M H, Kloker, D R and Dara-Abrams, B, 'Recovering parentheses from spoken algebraic expressions', IEEE Trans Audio and Electroacoustic, AV-21, pp 217-220, 1973.

Ostreicher, H J and Sharf, D J, 'Effects of coarticulation on the identification of detected consonant and vowel sounds', J Phonet, 4, pp 285-301, 1976.

Peterson, G E and Barney, H L, 'Control methods used in the study of

vowels', J Acoust Soc Amer, 24, pp 175-184, 1952.

Pfeiffer, L L, 'Technical description of the interactive laboratory system', Signal Technology Inc, Santa Barbara, 1978.

Pierce, J R, 'Whither speech recognition', J Acoust Soc Amer, 46, pp 1049-1051, 1969.

Pierce, J R, 'Whither speech recognition - II', J Acoust Soc Amer 47, pp 1616-1617, 1970.

Pisoni, D B, 'On the nature of categorical perception of speech sounds', PhD thesis, University of Michigan, 1971.

Pisoni, D B, 'Auditory and phonetic memory codes in the discrimination of consonants and vowels', Perception and Psychophysics, 13, pp 253-260, 1973.

Pols, L C W, van der Kamp, L J Th and Plomp, R, 'Perceptual and physical space of vowel sounds', J Acoust Soc of Amer, 46, pp 458-467, 1969.

Rothman, H B, 'A spectrographic investigation of consonant-vowel transitions in the speech of deaf adults', J Phonet, 4, pp 129-136, 1976.

Rubin, P, Turvey, M T and Van Gelder, P, 'Initial phonemes are detected faster in spoken words than in spoken nonwords', Perception and Psychophysics, 19, pp 394-398, 1976.

Samuel, A L, 'Whither speech recognition - A rebuttal', J Acoust Soc Amer, 47, p 1616, 1970.

Savin, H B and Bever, T G, 'The nonperceptual reality of the phoneme', J

Verbal Learning and Verbal Behaviour, 9, pp 295-302, 1970.

Sawsuch, J R and Pisoni, D B, 'On the identification of place and voicing features in synthetic stop consonants', J Phonet, 2, pp 181-194, 1974.

Searle, C L, Jacobson, J Z and Rayment, S G, 'Stop consonant discrimination based on human audition', J Acoust Soc Amer, 65, 3, pp 799-809, 1979.

Sharf, D J and Hemeyer, T, 'Identification of place of consonant articulation from vowel and formant transitions', J Acoust Soc Amer, 51, 2, 2, pp 652-658, 1972.

Stevens, K N, 'On the relations between speech movements and speech perception', Zeitschrift fur Phonetik, Sprachwissenschaft und Kommunikations Forschung, 21, pp 102-106, 1968.

Stevens, K N and Blumstein, S E, 'Invariant cues for place of articulation in stop consonants', J Acoust Soc Amer, 64, 5, pp 1358-1368, 1978.

Stevens, K N, and House, A S, 'Perturbations of vowel articulations by consonantal context: An acoustical study', J Speech Hearing Res, 6, pp 111-178, 1963.

Strange, W, Verbrugge, R R, Shankweiler, D P and Edman, T R, 'Consonant environment specifies vowel identity', J Acoust Soc Amer, 60, pp 213-224, 1976.

Streeter, L A, 'Language perception of 2 month old infants shows effects of both innate mechanisms and experience', Nature, 259, pp 39-41, 1976.

Streeter, L A and Nigro, G N, 'The role of medial consonant transitions in word perception', J Acoust Soc Amer, 65, 6, pp 1533-1541, 1979.

Strevens, P, 'Spectra of fricative noise in human speech', Language and Speech, 3, pp 32-49, 1960.

Su, L-S, Daniloff, R and Hammarberg, R, 'Variation in lingual coarticulation at certain juncture boundaries', Phonetica, 32, pp 254-263, 1975.

Umeda, N, 'Vowel duration in American English', J Acoust Soc Amer, 58, pp 434-445, 1975.

Wagner, M, 'Speaker Characteristics in Continuous Speech', PhD Thesis, Aust Nat Univ, Canberra, 1978.

Warren, R M, 'Perceptual restoration of missing speech sounds', Science, 167, pp 392-393, 1970.

Weinstein, C J, McCandless, S, Mondshein, L F and Zue, V W, 'A system for acoustic-phonetic analysis of continuous speech', IEEE Trans Acoust Speech Sig Proc, Vol ASSP-23, No 1, pp 54-72, 1975.

Weizenbaum, J, 'Contextual understanding by computers', in 'Recognising Patterns', P A Kolers and M Eden, eds, the MIT Press, Cambridge, Massachusetts, pp 170-193, 1968.

Wickelgren, W A, 'Context-sensitive coding, associative memory and serial order in (speech) behaviour', Psychological Review, 76, pp 1-15, 1969.

Wintz, H, LaRiviere, C and Herriman, E, 'Comments on the summarisation of the findings of Liberman et al regarding the role of formant transitions in the perception of voiceless stops', J Acoust Soc Amer, 58, pp 1333, 1975.

Wintz, H, Scheib, M E and Reeds, J A, 'Identification of stops and vowels

for the burst portion of /p,t,k/ isolated from conversational speech', J Acoust Soc Amer, 51, pp 1309-1317, 1972.

Woods, W A (principal author), 'Speech understanding system final report', BBN Report No 3438, November 1974-October 1976.

Zadeh, L A, 'Fuzzy sets', Information and Control, 8, pp 338-353, 1965.

Zadeh, L A, 'A fuzzy algorithmic approach to the definition of complex or imprecise concepts', Int J Man-Machine Studies, 8, pp 249-291, 1976.

Zlatin, M A and Koenigsknecht, R A, 'Development of the voicing contrast: Perception of stop consonants', J of Speech and Hearing Research, 18, pp 541-553, 1975.

Zwicker, E, Terhardt, E and Paulus, E. 'Automatic speech recognition using psychoacoustic models', J Acoust Soc Amer, 65, pp 487-498, 1979.

APPENDIX A

LIQUID-NASAL DIAGRAMS

Enlarged versions of the figures of Chapter 3 are given in this appendix. The numbering of Chapter 3 is retained throughout.

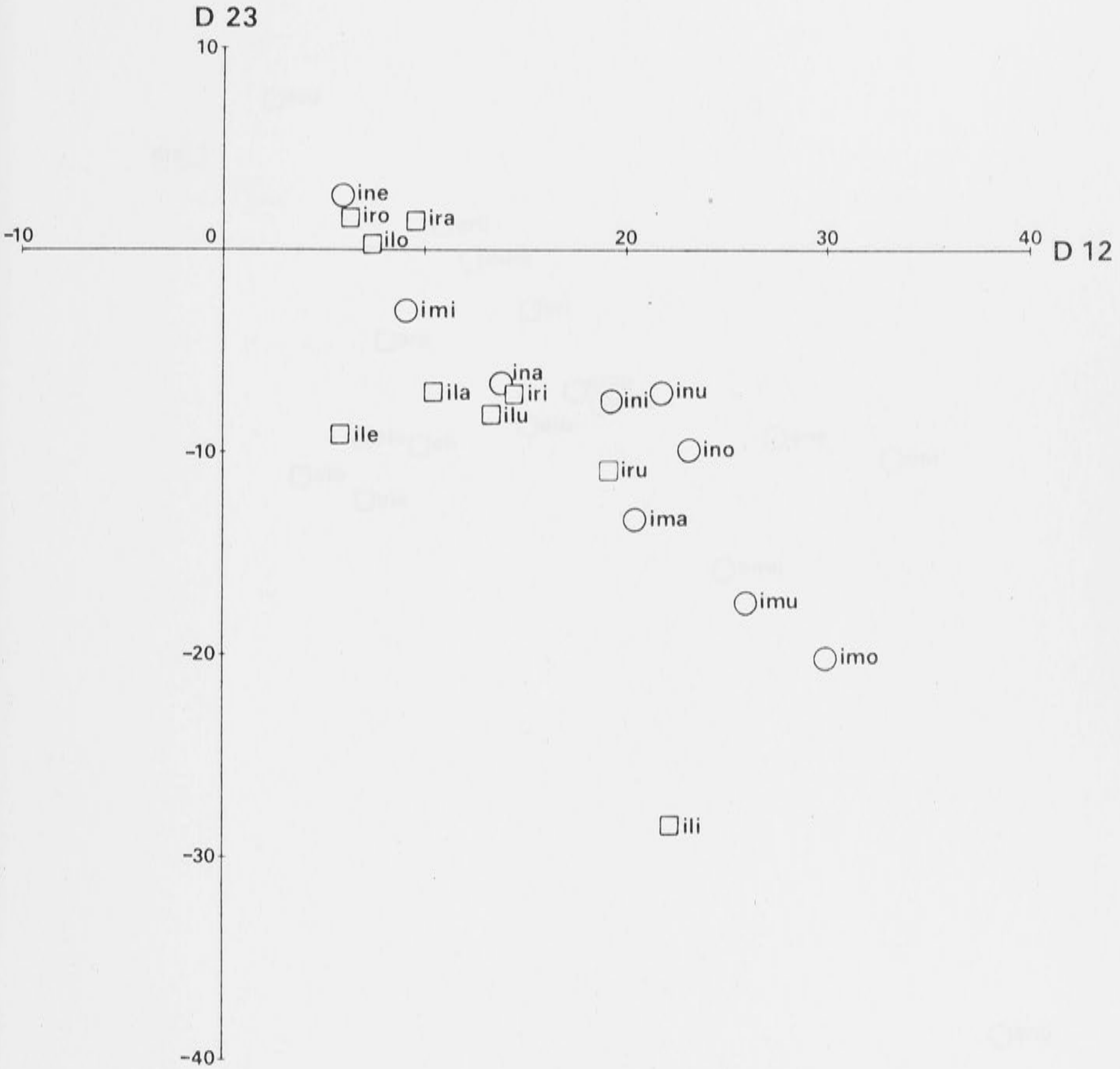


FIGURE 3.1(a): D23 versus D12 for the case /i/-consonant/-varying vowel.
Data for one speaker

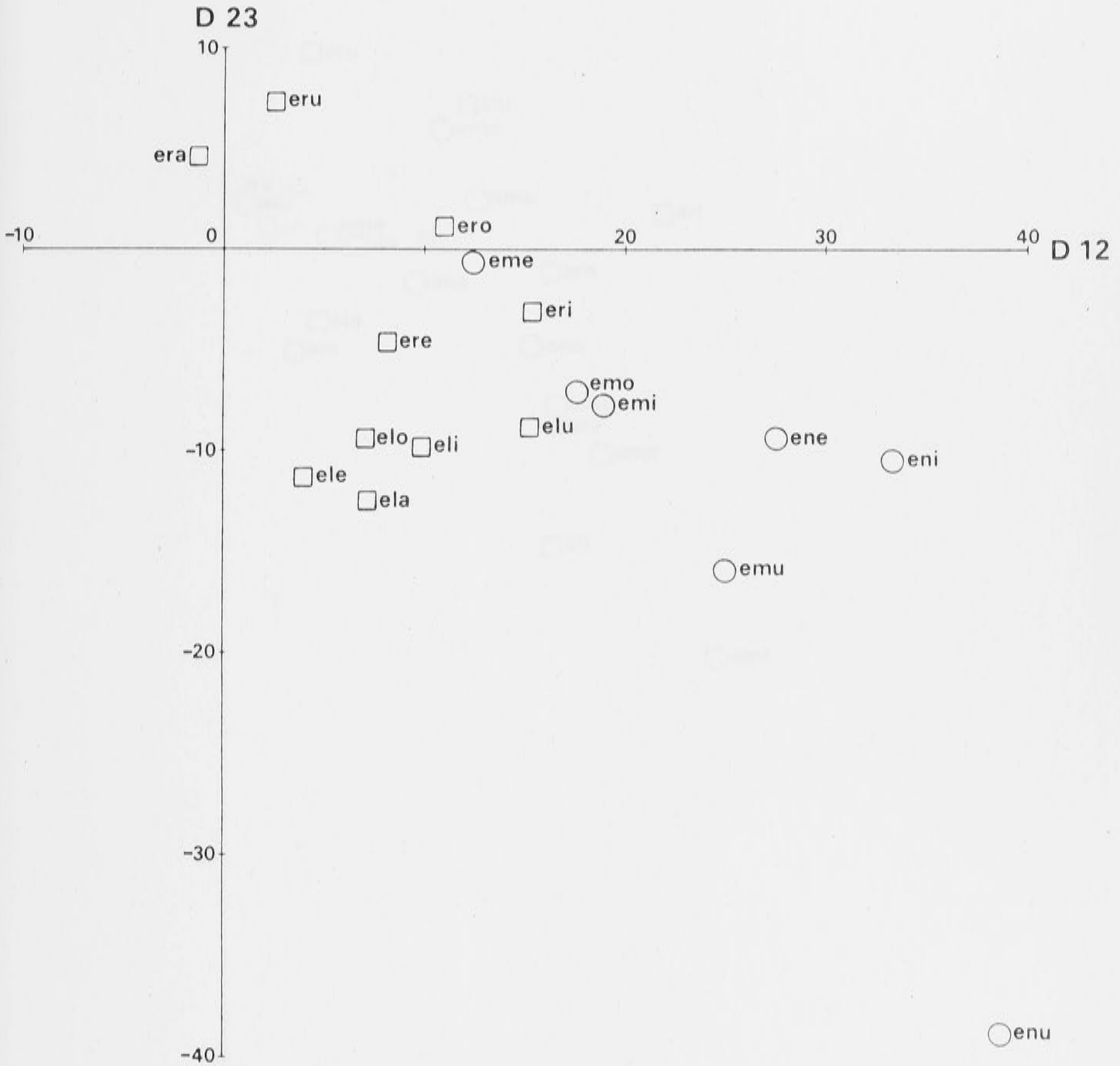


FIGURE 3.1(b): D23 versus D12 for the case /e/-consonant-varying vowel.
Data for one speaker.

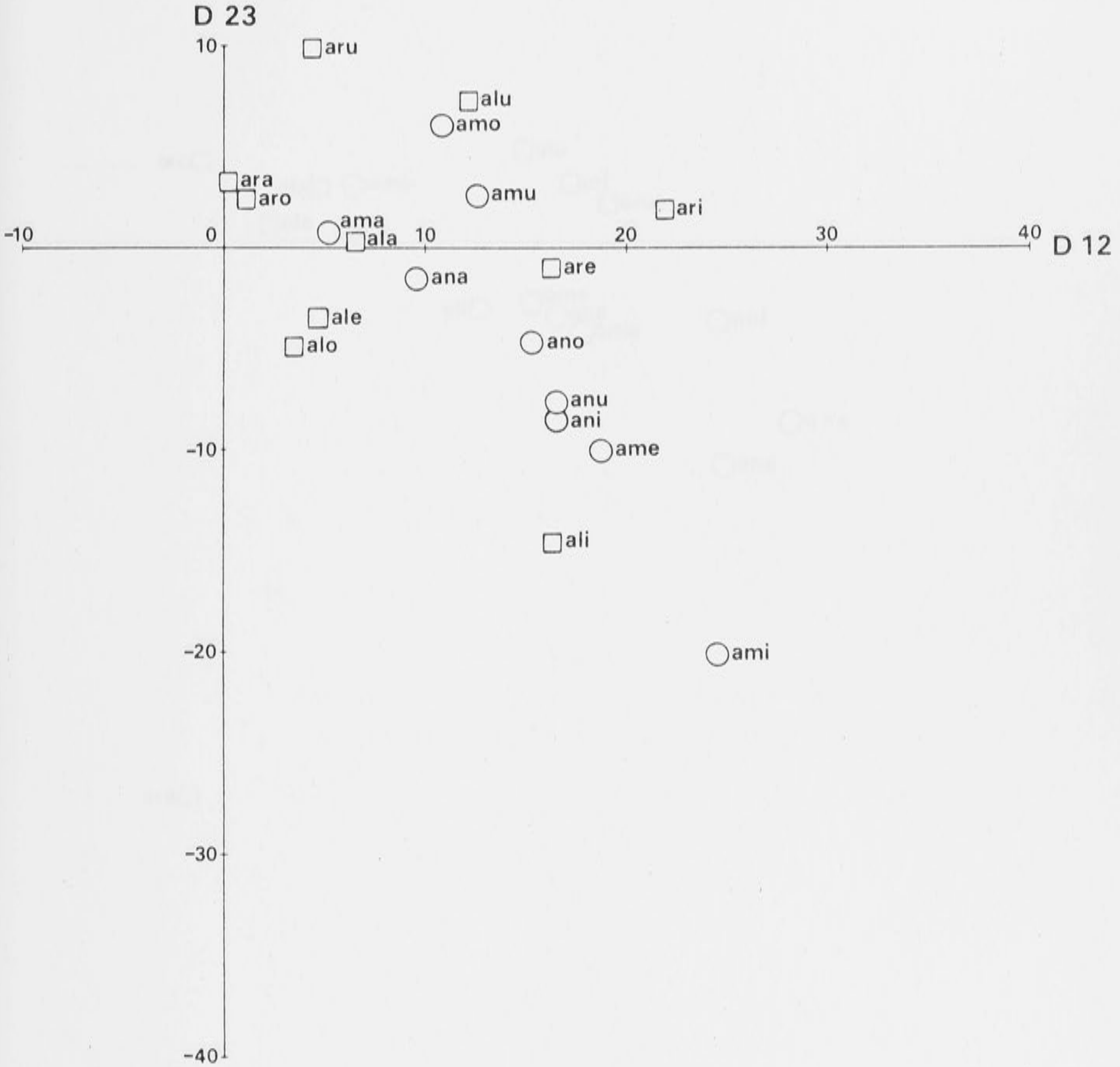


FIGURE 3.1(c): D23 versus D12 for the case /a/-consonant-varying vowel.
Data for one speaker.

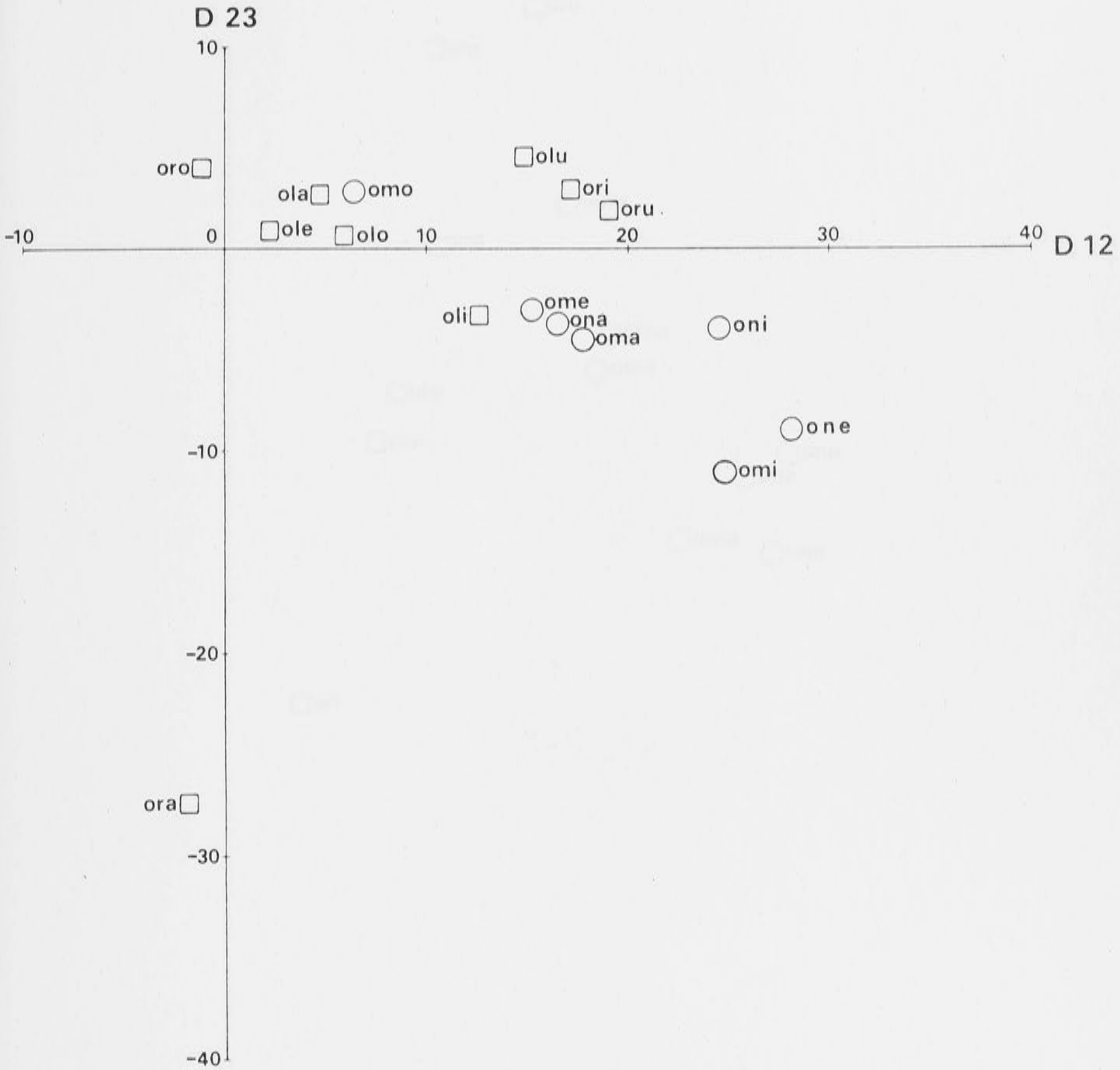


FIGURE 3.1(d): D23 versus D12 for the case /o/-consonant-varying vowel.
Data for one speaker.

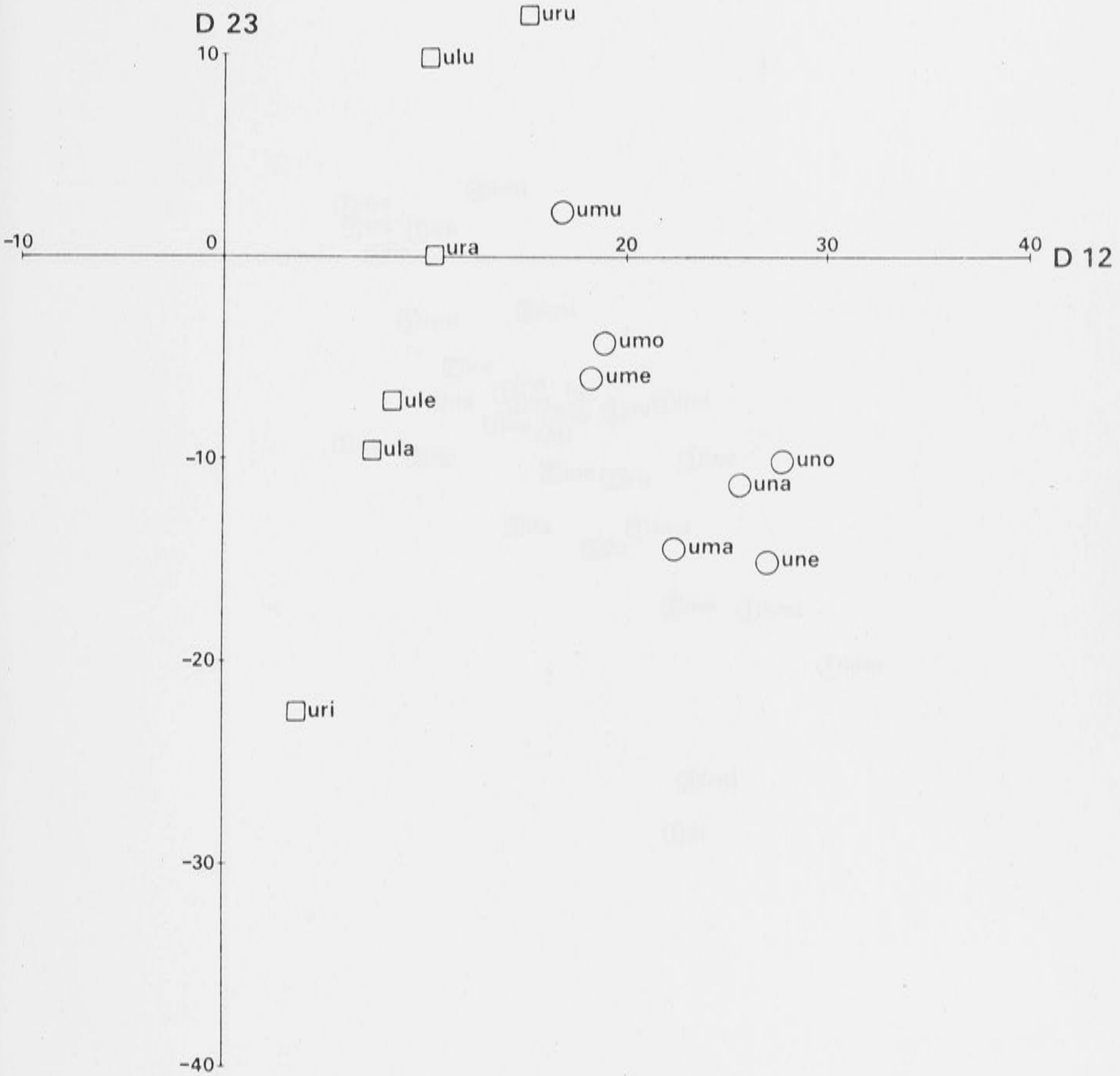


FIGURE 3.1(e): D23 versus D12 for the case /u/-consonant-varying vowel.
Data for one speaker.

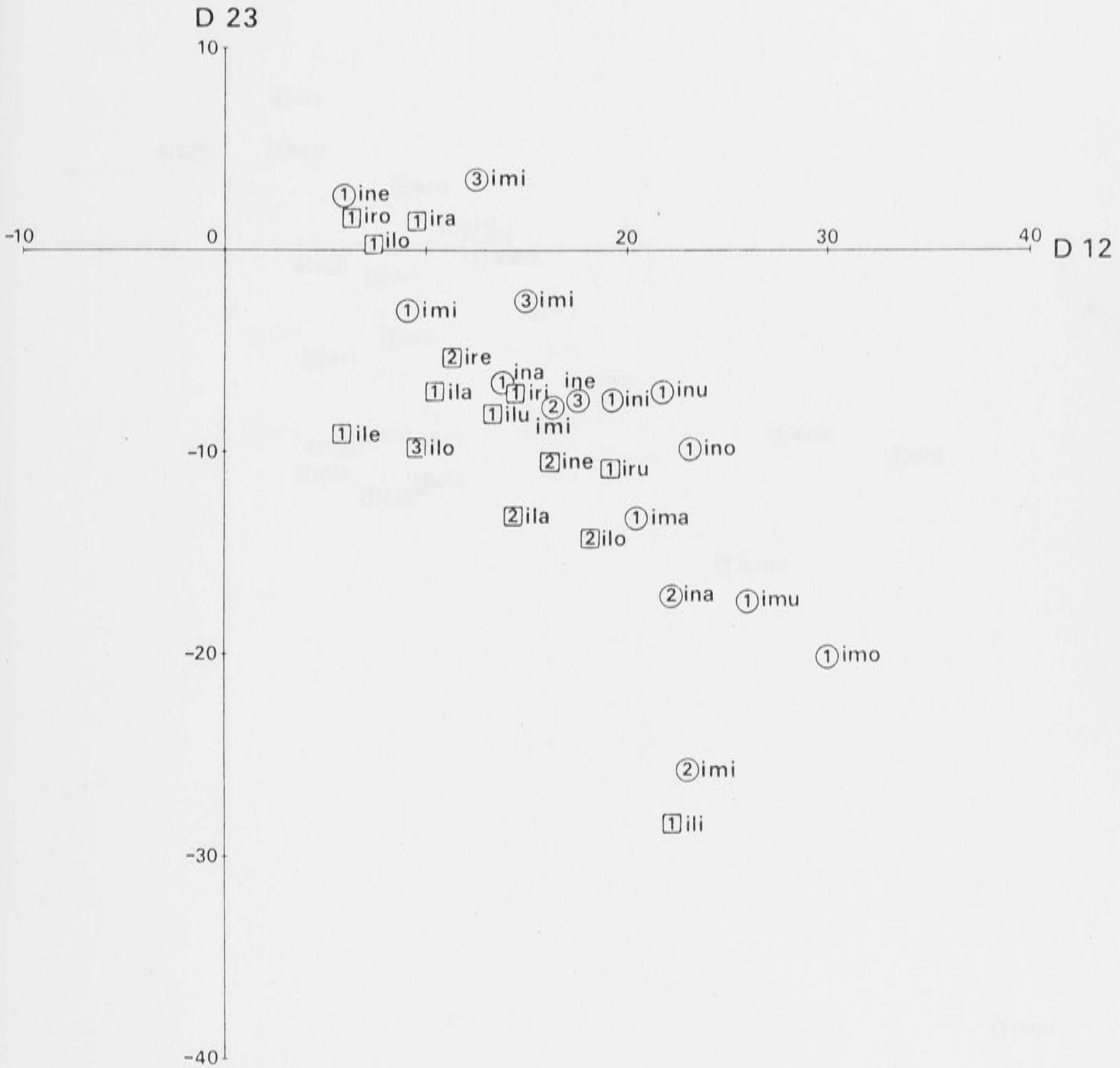


FIGURE 3.2(a): D23 versus D12 for the case /i/-consonant-varying vowel.
Data for four speakers.

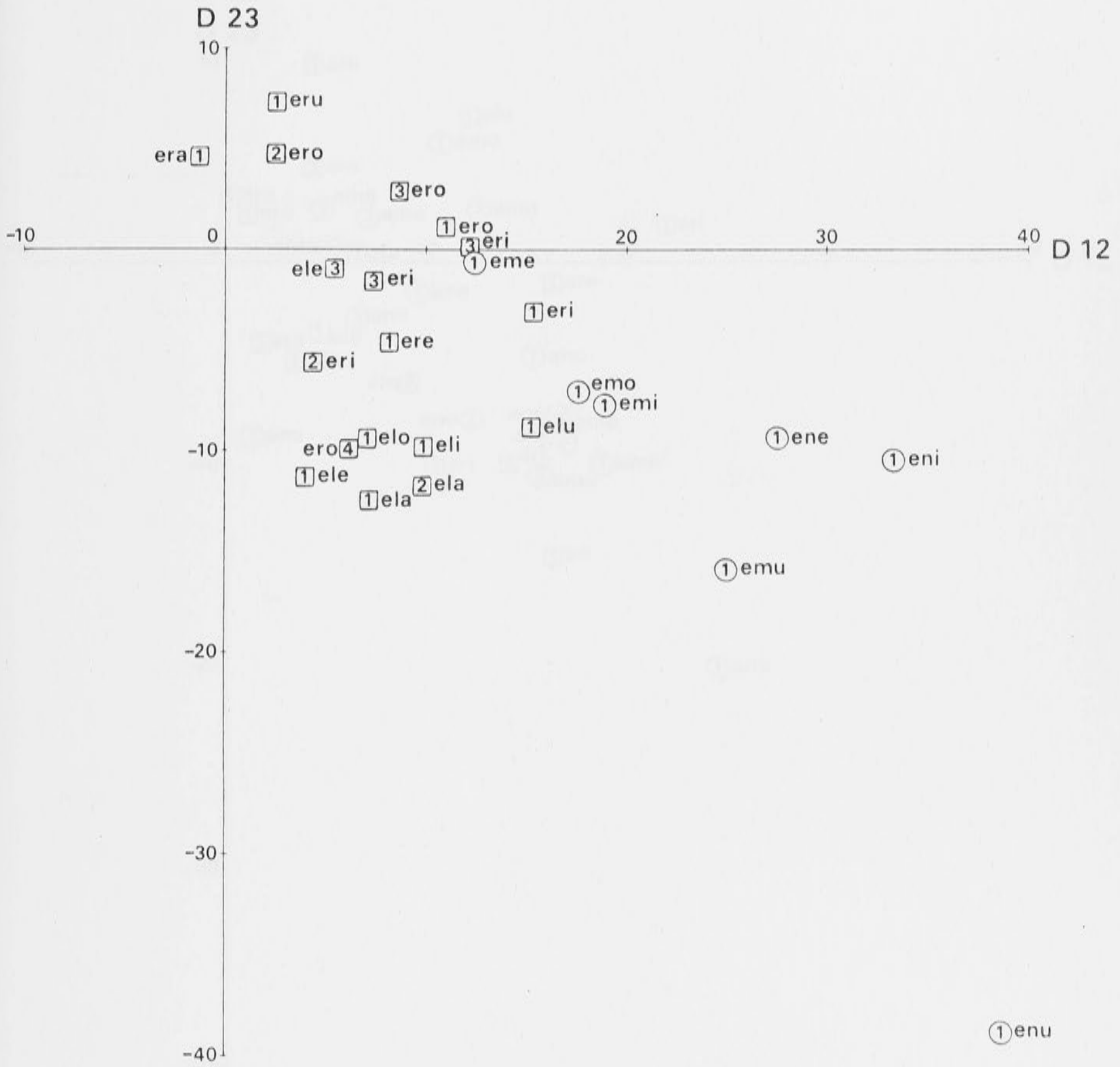


FIGURE 3.2(b): D23 versus D12 for the case /ė/-consonant-varying vowel.
Data for four speakers.

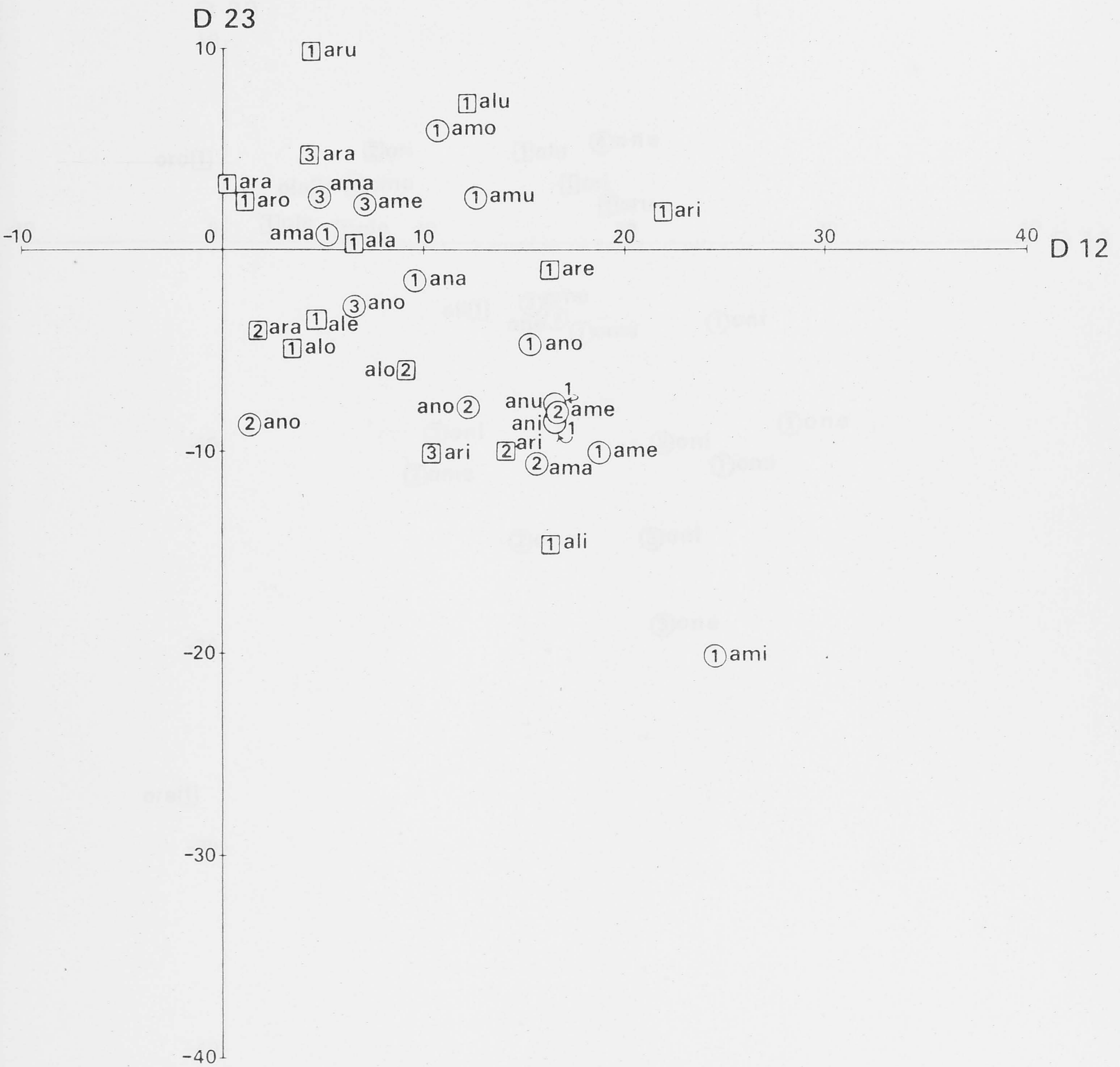


FIGURE 3.2(c): D23 versus D12 for the case /a/-consonant-varying vowel.
Data for four speakers.

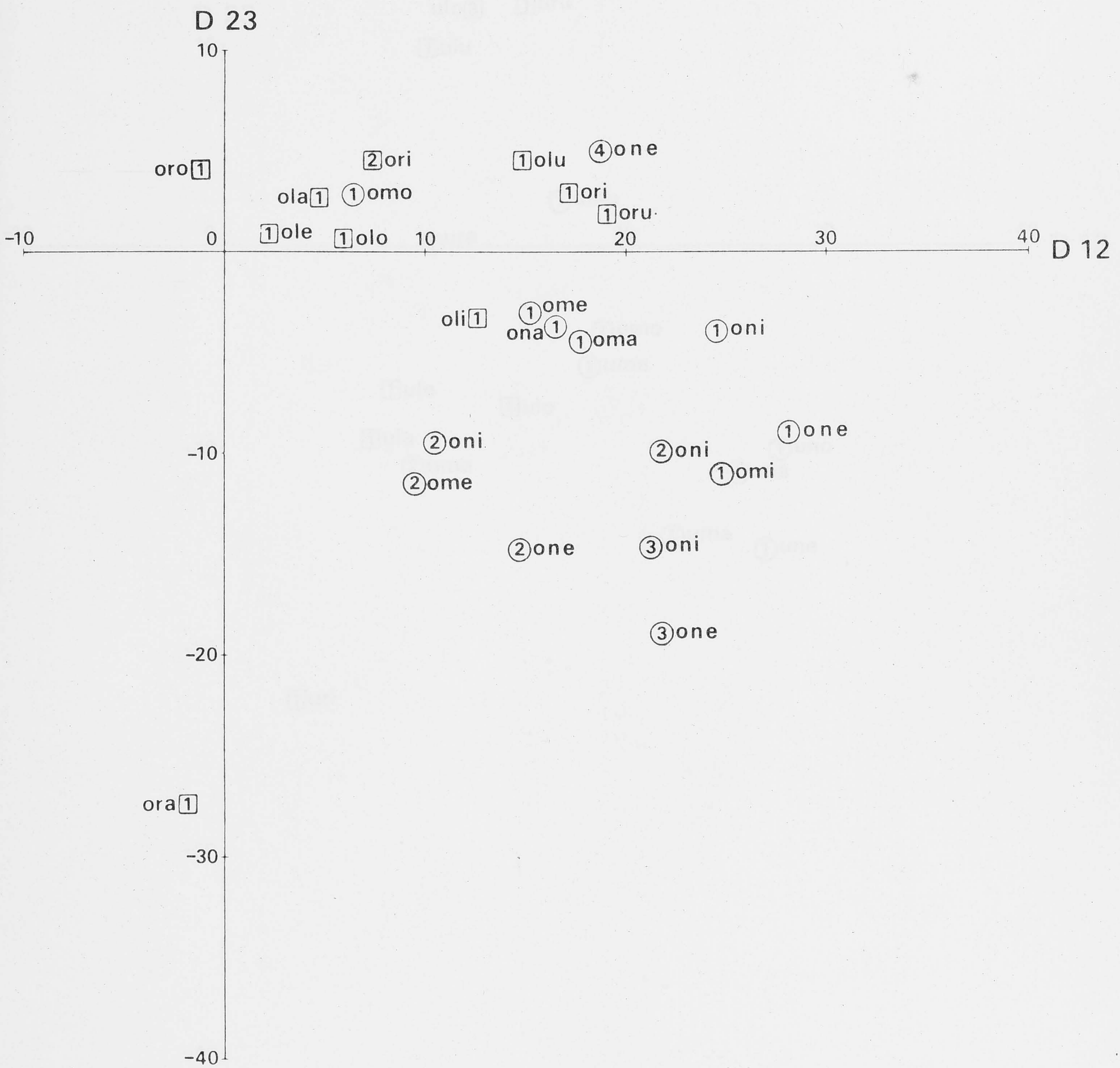


FIGURE 3.2(d): D23 versus D12 for the case /o/-consonant-varying vowel.
Data for four speakers.

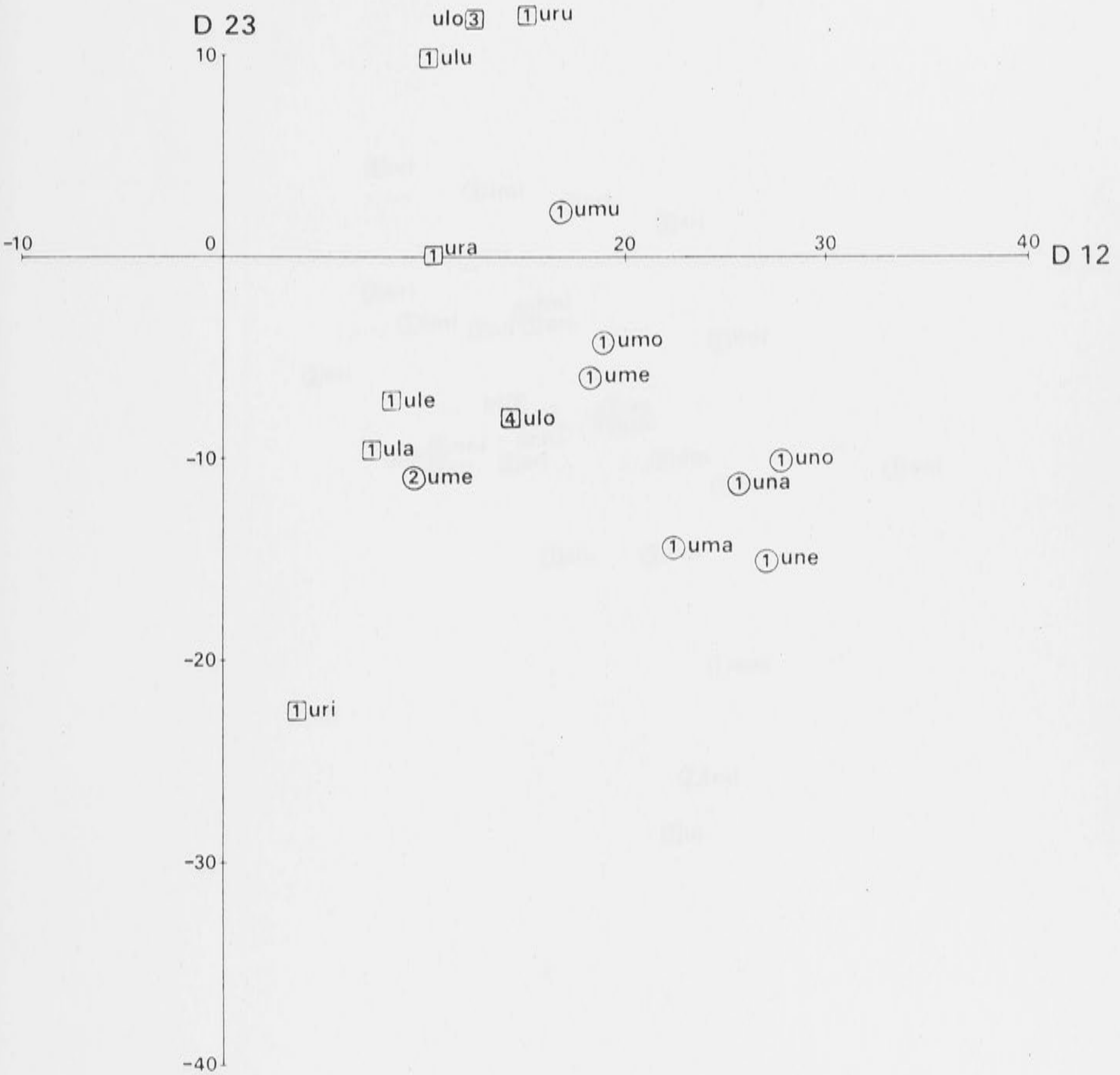


FIGURE 3.2(e): D23 versus D12 for the case /u/-consonant-varying vowel.
Data for four speakers.

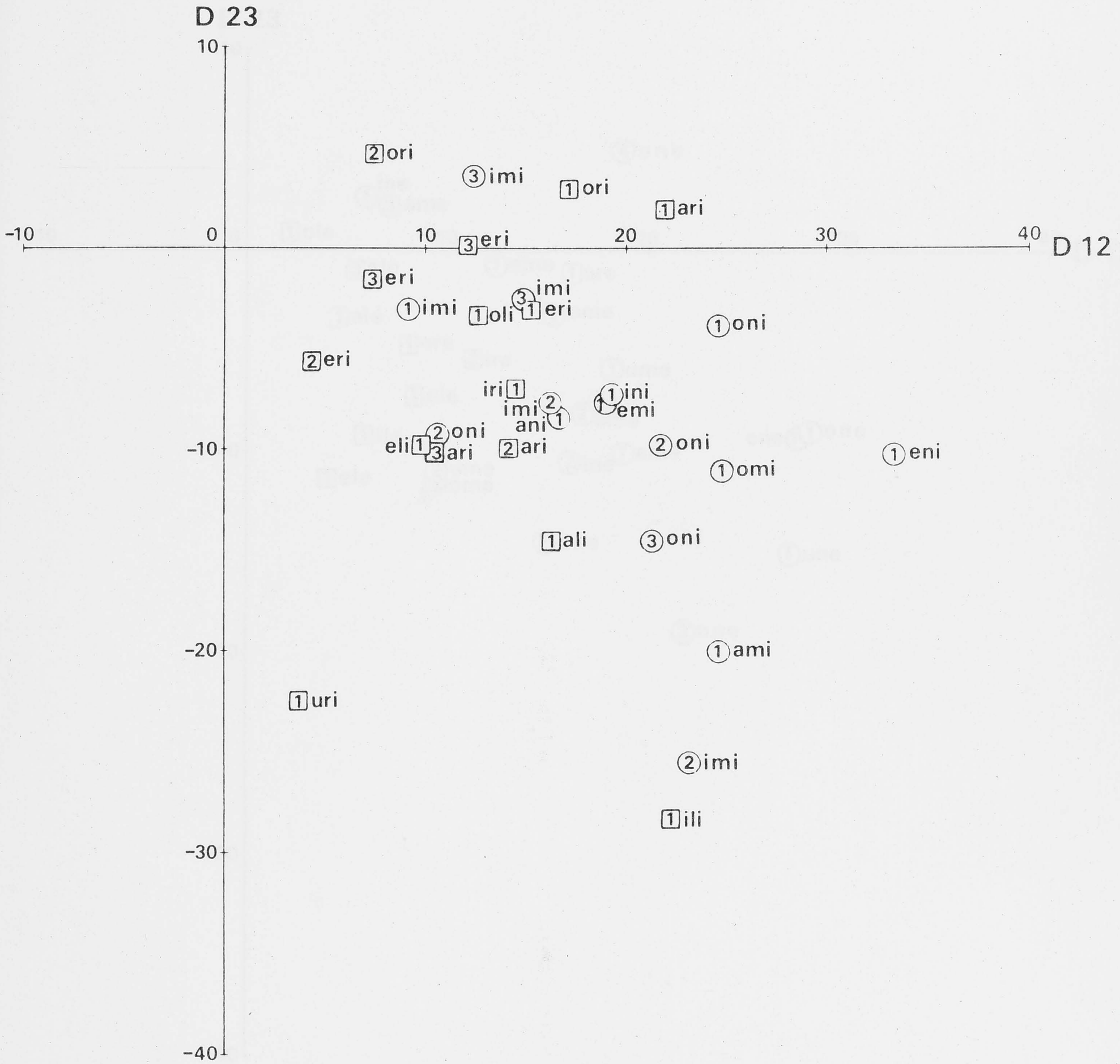


FIGURE 3.3(a): D23 versus D12 for the case varying vowel-consonant-/i/. Data for four speakers.

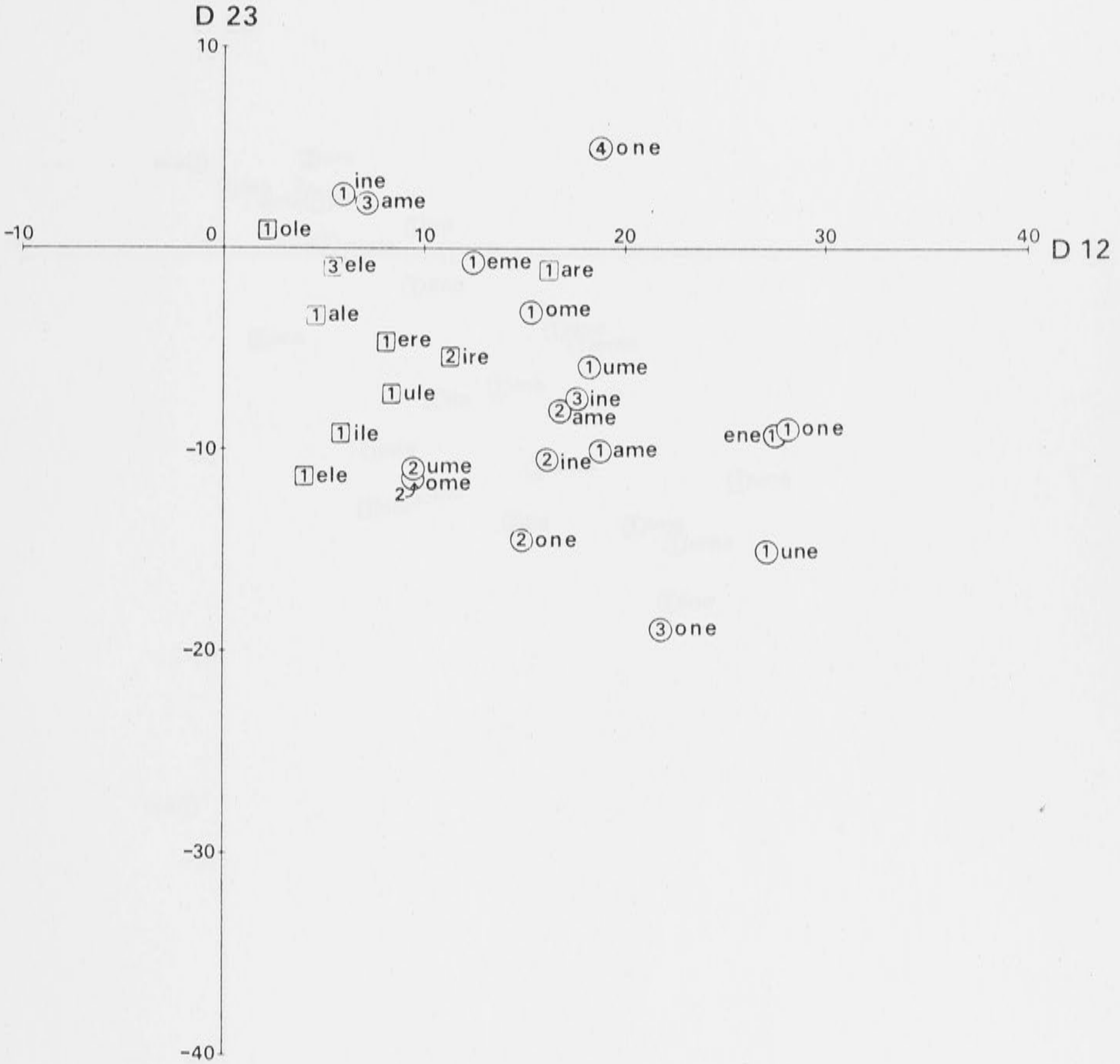


FIGURE 3.3(b): D23 versus D12 for the case varying vowel-consonant-/e/. Data for four speakers.

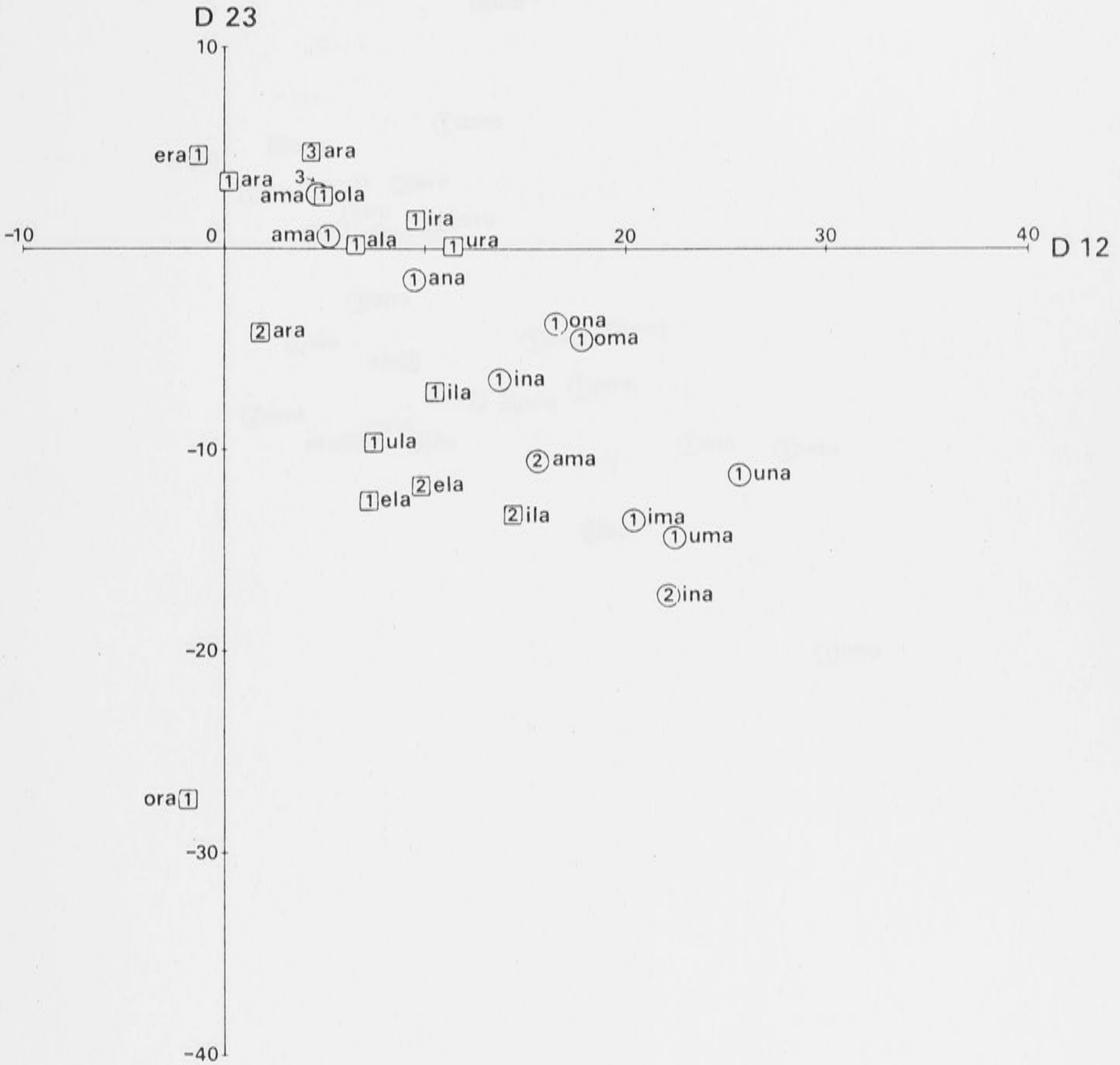


FIGURE 3.3(c): D23 versus D12 for the case varying vowel-consonant-/a/. Data for four speakers.

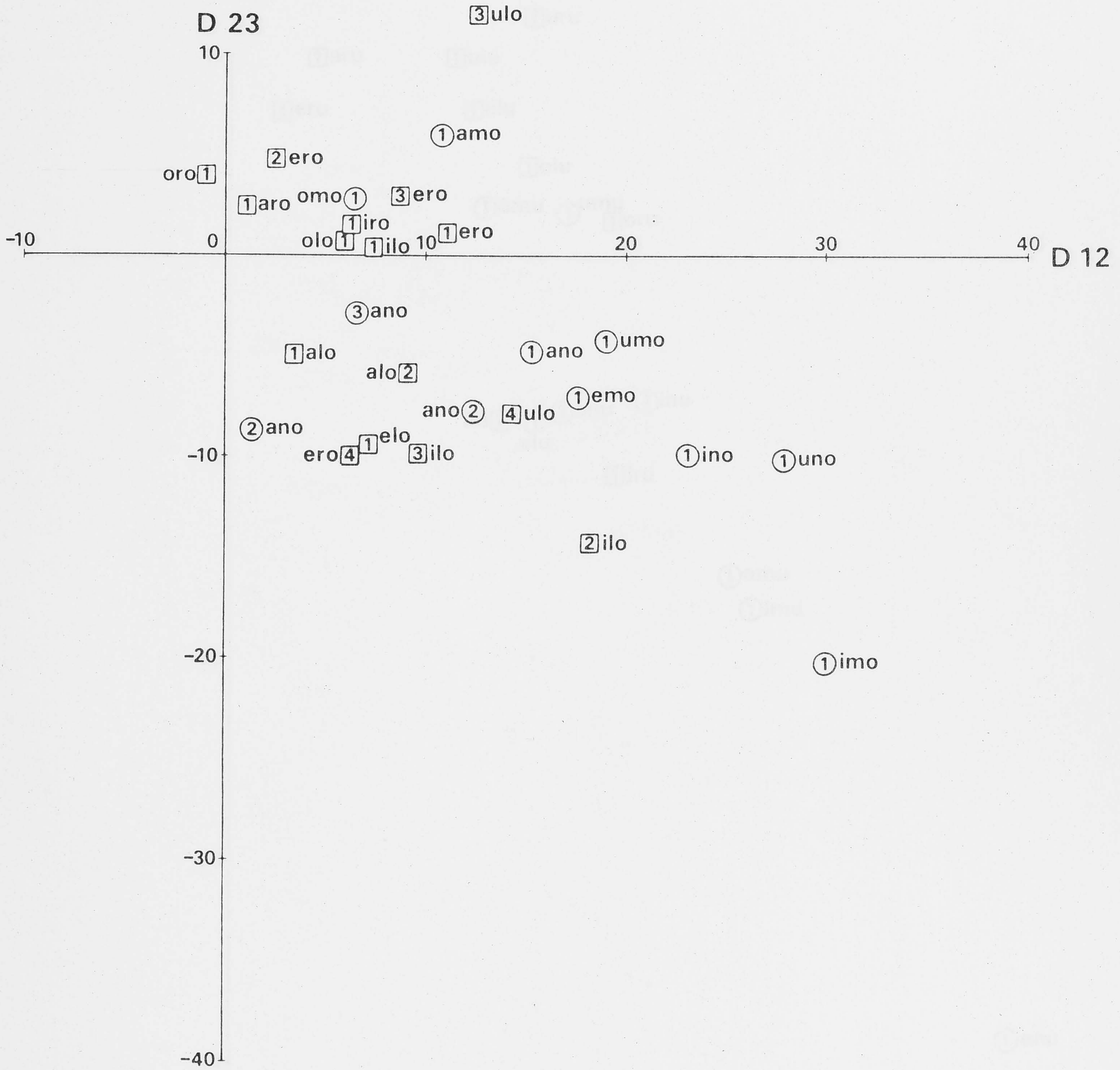


FIGURE 3.3(d): D23 versus D12 for the case varying vowel-consonant-/o/. Data for four speakers.

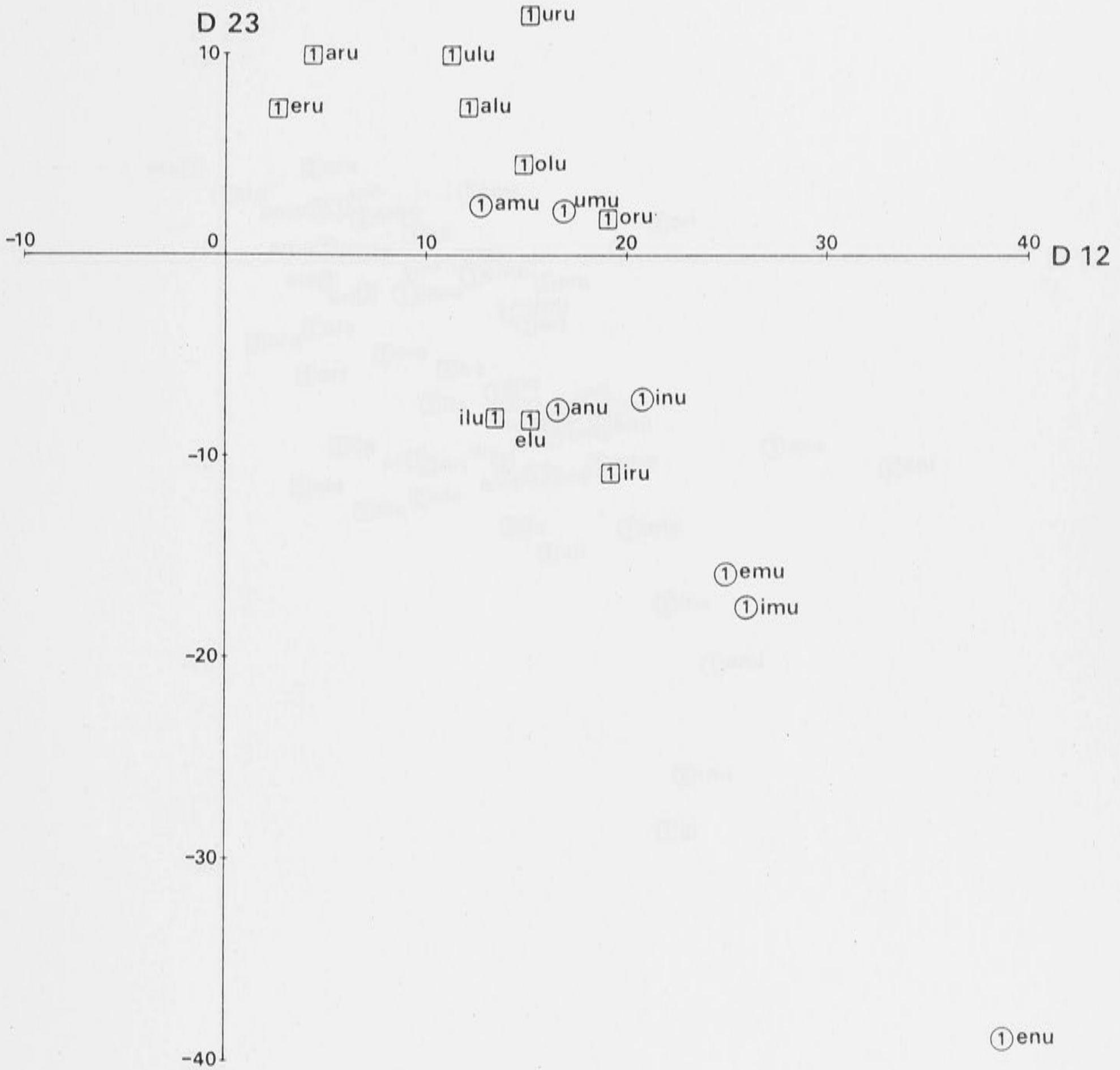


FIGURE 3.3(e): D23 versus D12 for the case varying vowel-consonant-/u/. Data for four speakers.

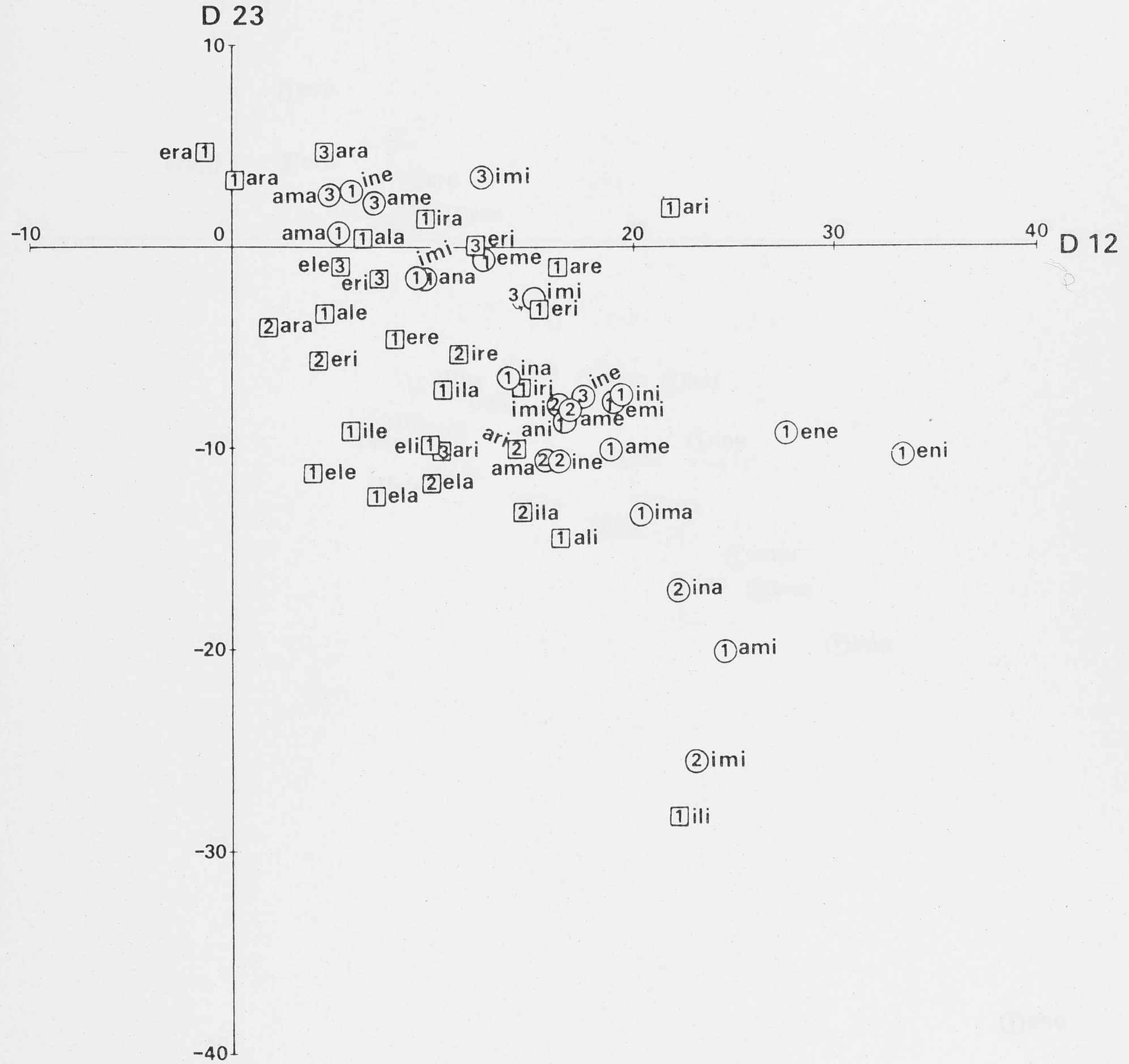


FIGURE 3.4(a): D23 versus D12 for the case (front or central vowel)-consonant-(front or central vowel).
Data for four speakers.

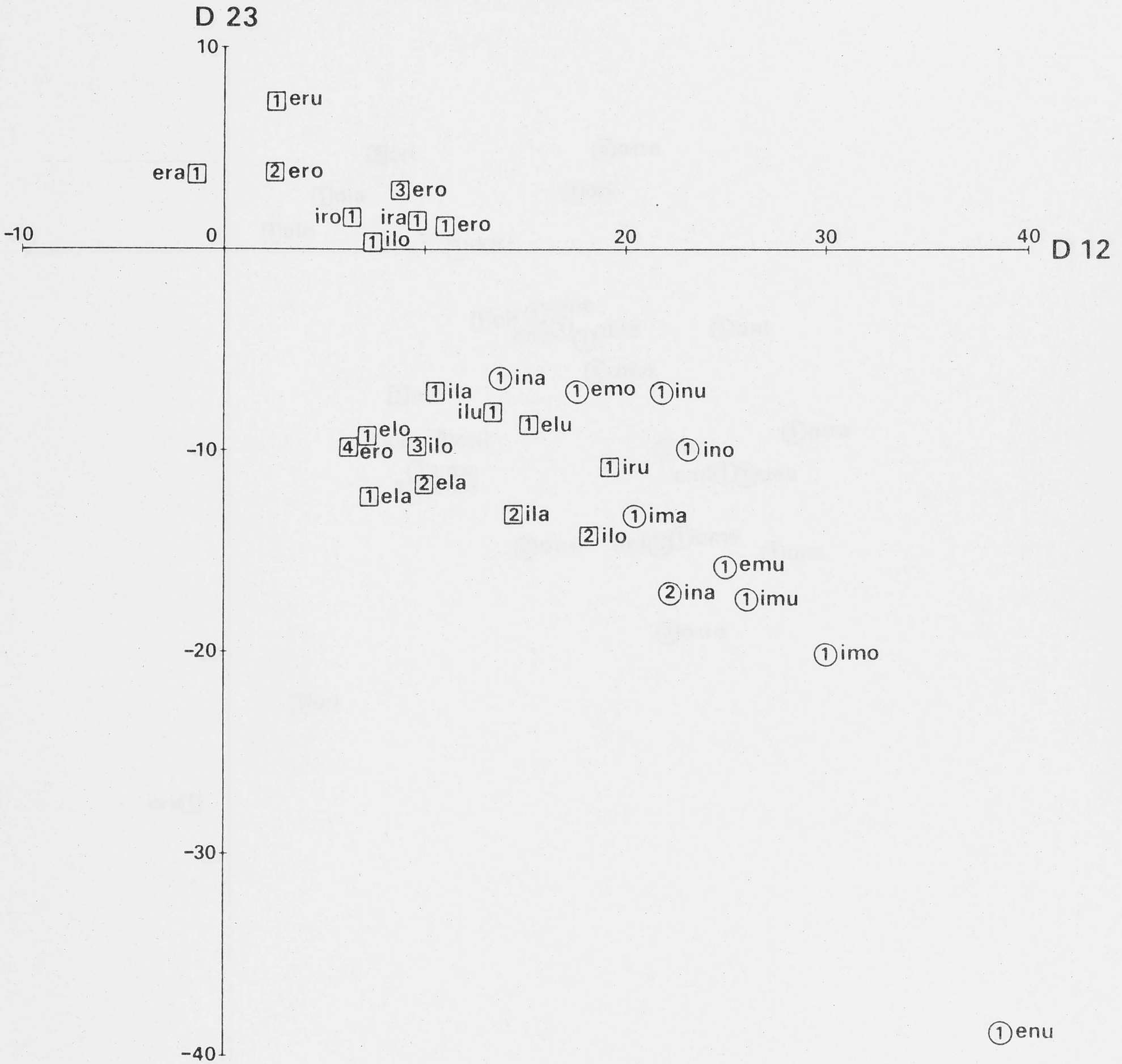


FIGURE 3.4(b): D23 versus D12 for the case (front vowel)-consonant-(central or back vowel).
Data for four speakers.

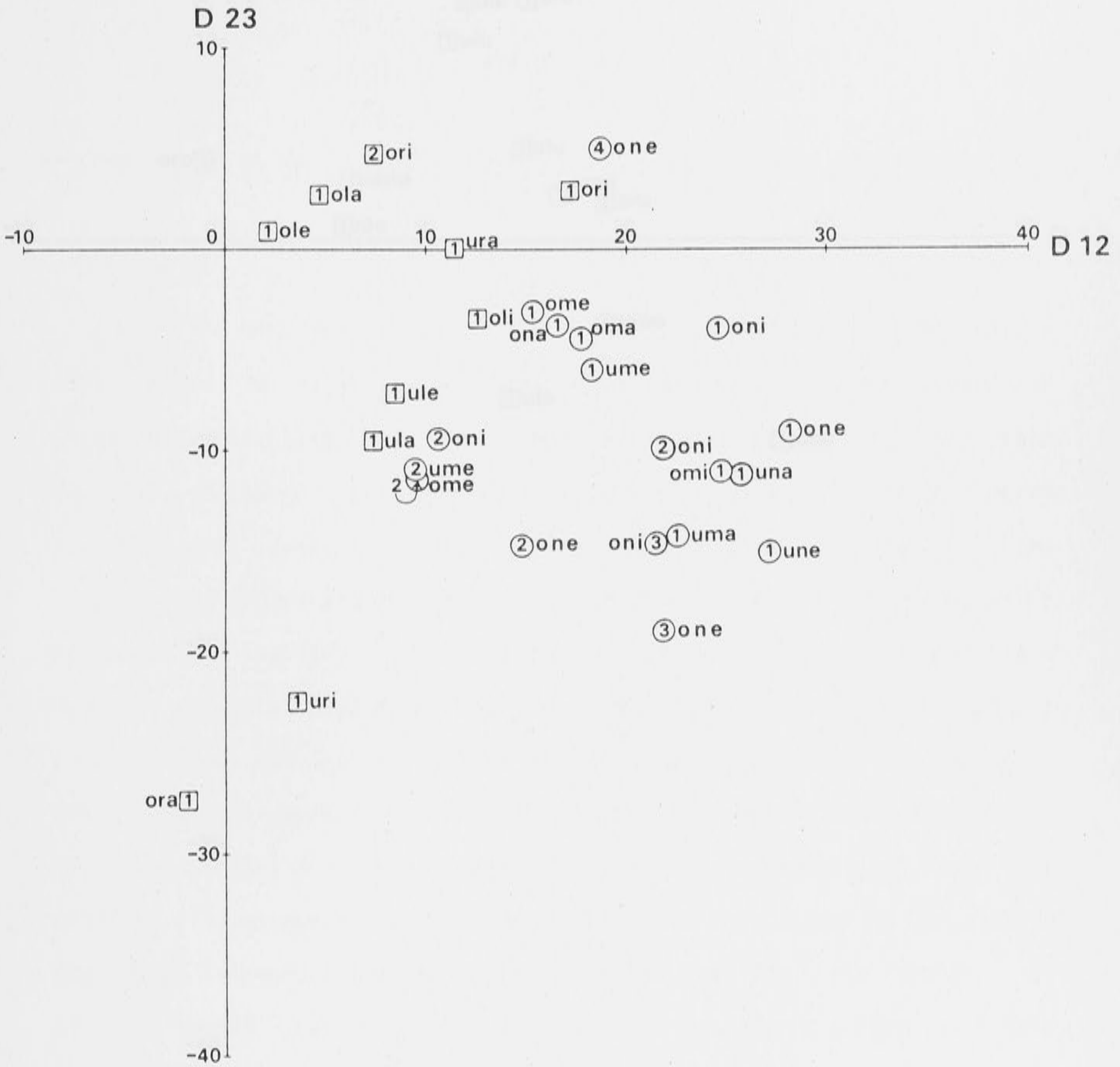


FIGURE 3.4(c): D23 versus D12 for the case (back vowel)-consonant-(front or central vowel).
Data for four speakers.

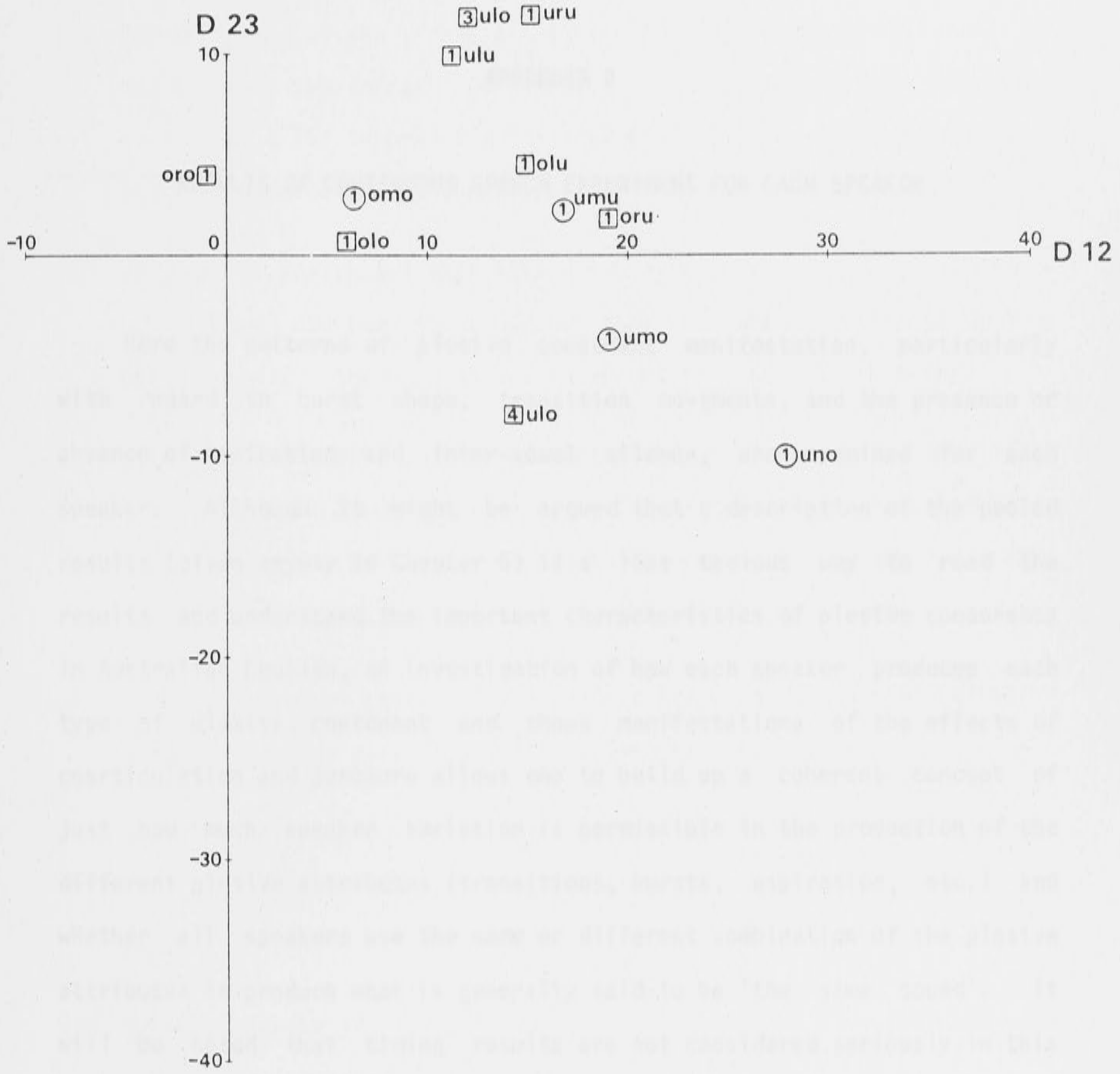


FIGURE 3.4(d): D23 versus D12 for the case (back vowel)-consonant-(back vowel).
Data for four speakers.

APPENDIX B

RESULTS OF CONTINUOUS SPEECH EXPERIMENT FOR EACH SPEAKER

Here the patterns of plosive consonant manifestation, particularly with regard to burst shape, transition movements, and the presence or absence of aspiration and inter-vowel silence, are examined for each speaker. Although it might be argued that a description of the pooled results (given anyway in Chapter 6) is a less tedious way to read the results and understand the important characteristics of plosive consonants in Australian English, an investigation of how each speaker produces each type of plosive consonant and shows manifestations of the effects of coarticulation and juncture allows one to build up a coherent concept of just how much speaker variation is permissible in the production of the different plosive attributes (transitions, bursts, aspiration, etc.) and whether all speakers use the same or different combination of the plosive attributes to produce what is generally said to be 'the same sound'. It will be noted that timing results are not considered seriously in this Appendix. This is so because the timing measures investigated can be described simply in a quantitative fashion (See Section 6.6.3). Here remarks on aspects of plosive timing are confined to cases where timing effects interact with other attributes of the consonant under discussion.*

* It should be noted that phonetic symbols are in Letter Gothic type in this appendix.

For speaker 1:

For the cases when the consonant in the VCV combination is bilabial:

(1) For the cases of /Vb V/ the burst is either barely noticeable; or it is a burst-cum-formant structure; or it is nonexistent. In other words when a /b/ is word-final the burst is weak. In these cases there is also never any aspiration or silence between the vowels.

(2) /b/ and /p/ bursts are generally diffuse in structure (i.e. standard b, p shape) the hump being approximately 1.5 kHz in width and tailing off gradually. The hump generally occurs between 1.5 and 4 kHz.

(3) Bilabial burst coarticulatory effects are not very marked although for /iCi/ cases the burst is generally in the 2.5-4 kHz region; the two /)C V/ cases the burst hump occurs in the region 0-1.5kHz. 0-1.5 kHz. For all other cases it ranges, as mentioned from 1.5-4 kHz.

(4) There are several cases of peakiness in the bilabial bursts; either a sudden mid-range peak or peakiness-over-the-diffuseness.

(5) Aspiration is not generally noticeable in the voiced bilabial bursts but is apparent in the unvoiced bursts, particularly in the /V CV/ cases.

(6) Silence between vowels is found in all-but-one /V CV/ case and in all-but-one unvoiced /VC V/ case.

(7) There are generally marked downwards F2 transitions when one of the vowels is /i/.

(8) F2 /)/ transitions are almost straight.

(9) F3 transitions are either not noticeable or are straight.

(10) F1 transitions are sometimes downwards and sometimes straight.

(11) Juncture effects are not consistently noticeable in the transitions although coarticulation effects are noticeable.

For the cases when the consonant is alveolar:

(1) Alveolar bursts are strong for this speaker; the hump is varying in width but is generally about 1.5 kHz in width and is positioned between 1.5 and 5 kHz. By a 'strong alveolar burst' it is meant that the drop-off from the hump is rapid and steep on both sides.

(2) There is a tendency for unvoiced bursts to be higher in frequency than the voiced bursts.

(3) In both voiced and unvoiced /VC V/ cases there is often a burst-cum-formant effect. This means that the release spectrum displays not only a burst but also vestiges of the formants, particularly the first formant.

(4) Coarticulatory effects or the burst are most noticeable when comparing /iCi/ and /)C)/ bursts. For the former case has the hump is generally at a higher frequency than the latter. When one vowel is /i/ and the other /)/ no particular effect is evident.

(5) Aspiration occurs in all /t/ and in no /d/ cases.

(6) There is no silence between the vowels in the voiced cases and the only noticeable silence occurs in the unvoiced /V CV/ cases.

(7) There are generally well marked transitions. For cases where the vowel is /i/ the F2 points down slightly (but not as much as in the bilabial case); for transitions associated with /)/ F2 points up. F1

generally points down.

(8) F3 is not marked. When there is some transition noticeable it generally tends to point down.

(9) There is a strongly marked symmetrical transition pattern in the /)C)/ cases. Both F1 transitions point down, both F2 transitions point up, and the F3 transitions are straight or slightly down.

For the cases when the consonant is velar:

(1) The typical velar burst (i.e. two-peak) is very evident for this speaker. For the voiced /VC V/ case there is always a burst-cum-formant structure; it is also seen occasionally in the voiceless case. The first peak occurs in the range 1.1-2.5 kHz; and the second peak in the range 3.6-4.2 kHz.

(2) Coarticulation effects are strong, in the expected directions. A tendency for the burst to show strongest coarticulation with the V on the same side of the consonant as the word boundary is in evidence in several cases.. But the effect of both vowels can be seen.

(3) There is aspiration (at least briefly) in all but one /V CV/ case but it only appears for two /VC V/ cases, both unvoiced.

(4) In all cases there is silence between the vowels in the unvoiced cases but not in the voiced cases.

(5) There is the very marked transition effect whereby when one of the vowels is /i/; the F2 and F3 transitions approach, i.e. F2 up and F3 down. This is particularly noticeable for /iCV/.

(6) There is evidence of coarticulation in that sometimes F2 transitions

from /)/ are straight and sometimes they go up. The hypothesis that the locus position associated with the vowel following the consonant is always dominant is by no means always supported in cases where both /i/ and /)/ are present.

(7) F1 transitions almost always point down in velar consonant cases.

For speaker 2:

For the cases when the consonant is bilabial:

(1) bilabial bursts are very diffuse indeed. They often look like a straight line. Humps are very slight and occur anywhere in the frequency range 0-4 kHz. Burst-cum-formant structures are common. Some peakiness is apparent.

(2) Bursts are more noticeable in the /V CV/ than in the /VC V/ cases for the voiced plosives.

(3) For both /ipi/ bursts there is a strange shape with a peak at 1.6 kHz.

(4) No signs of coarticulation in the burst structure.

(5) Aspiration is only consistently noticeable in the /V CV/ cases for unvoiced plosives.

(6) Silence is only ever found under the same conditions for aspiration as stated in (5).

(7) Transitions take the expected form. The most noticeable effect (particularly in contradistinction to speaker (1) is that the F3 transitions particularly as associated with the vowel /i/ are very marked and point downwards.

For the cases when the consonant is alveolar:

(1) Alveolar bursts are typical in shape and are very prominent particularly for the unvoiced case of /)t)/. The hump can vary in size and occurs anywhere in the range 1.5-5 kHz.

(2) Coarticulation effects on the bursts are marked, with /iCi/ being high in frequency (2.5-5 kHz) the bursts associated with /)t)/ being low in frequency. The /iC)/ bursts tend to be low (not as low as /)C)/ bursts) and the /)Ci/ bursts tend to be high (though not as high as for /iCi/). This last mentioned effect is independent of the juncture position.

(3) Burst-cum-formant structures appear to some extent in both /VC V/ and /V CV/ cases.

(4) Aspiration occurs with all the unvoiced /V CV/ cases but with only two of the unvoiced /V CV/ cases and essentially not at all with the voiced cases.

(5) Silence between vowels is seen in the unvoiced /V CV/ cases and practically nowhere else.

(6) In the /iCi/ cases transitions are noticeable or marked. For the /)C)/ cases the transitions are very marked with the characteristic F1-down, F2-up, F3-straight symmetrical pattern. For the cases involving /i/ or /)/ the F2 transition is generally very marked particularly for the transition between the consonant and the vowel which is on the same side of the word boundary as it is.

For the cases when the consonant is velar:

(1) Velar bursts are typical in shape (i.e. two-peak, or sometimes three-peak with both end peaks high and the middle one smaller). Sometimes peak 2 (assuming the burst is only two-peak) is higher than peak 1.

(2) For the /iCi/ case the peaks are close together and it seems that there is often a high region between the peaks which can make the burst look not unlike an alveolar burst.

(3) Coarticulation effects are noticeable particularly on the first peak, and there seems to be an interaction between coarticulation and juncture effects. Generally the following vowel can exert the major influence but the timing between the vowels can reverse this. Not surprisingly if the timing is not too small the coarticulation pattern seems to depend mainly on which vowel is on the same side of the word boundary as the consonant (e.g. /ik)/ shows major evidence of coarticulation with /i/).

(4) Aspiration is always associated with the unvoiced velar /k/ and often also with /g/.

(5) There is generally no silence between the vowels in the voiced cases. There is always silence in the unvoiced /V CV/ cases and sometimes in the /VC C/ cases.

(6) Sometimes the transitions associated with the unvoiced cases are very slight. The usual effect whereby F2 and F3 approach is seen in the /iCi/ cases. The locus seems to be associated with the preceding vowel when both /i/ and /)/ are involved.

For speaker 3:

For the cases when the consonant is bilabial:

(1) Highly idiosyncratic bilabial bursts. It must be remembered that this person is of German origin having spoken German exclusively until the age of twelve although his English is now so good that new acquaintances do not realize that it is his second language. The bursts are almost always present and well-marked, some particularly for the /VC V/ case are burst-cum-formant structures. There are several 'standard', very diffuse bilabial bursts in the range 1-4.5 kHz and with hump width of about 2 kHz on average. However there are several bursts, particularly unvoiced bursts which are very peaky and look somewhat like standard three-peak velar bursts.

(2) There is no real evidence of coarticulation effects in the bursts.

(3) Aspiration seems standard for unvoiced /V CV/ cases and occasionally appears in other situations.

(4) Silence is also present in unvoiced /V CV/ cases and also occasionally appears in other situations.

(5) Transitions particularly F2 transitions are well-marked especially when one of the vowels is /i/. This suggests a very low bilabial locus. Transitions in /V CV/ cases are sometimes confined to following transitions.

For the cases when the consonant is alveolar:

(1) /t/ bursts are higher in both amplitude and frequency position than

/d/ bursts.

(2) The burst hump can be anywhere in the range 1.2-5 kHz. Many bursts are very marked; some are diffuse but still recognizably alveolar bursts. There are several burst-cum-formant structures in the /VC V/ cases.

(3) There is no real evidence of coarticulation in the burst hump position.

(4) Aspiration is very common and is present in all /V CV/ cases and all but two /VC V/ cases.

(5) Silence is present in all unvoiced /V CV/ cases, in some voiced /V CV/ cases, and is essentially not present in any /VC V/ cases.

(6) Transitions are almost always present and are very well marked. In the /)C)/ cases the characteristic F1-down, F2-up, F3-straight pattern is very noticeable.

For the cases when the consonant is velar:

(1) The regular two-peak burst is the predominant shape. However for three /iCi/ cases and the /) gi/ case it seems as though the two peaks have merged to give a very large, very narrow peak at 2.7 kHz (and always exactly there). Perhaps however it is just that the lower-frequency peak 'swamps' the second peak because there is some evidence for a small peak at 4.3 kHz.

(2) For the two /Vg)/ cases there is no burst.

(3) Coarticulation effects in the bursts indicate that the following vowel has the predominant effect except for the case of /)k i/; this

exception is not correlated with a large time-gap.

(4) Aspiration occurs in all the /V CV/ cases and all the unvoiced /VC V/ cases but in none of the voiced /VC V/ cases.

(5) Silence appears in most of the /V CV/ cases but only rarely (and then for examples where the consonant is unvoiced) in the /VC V/ cases.

(6) For /iCi/ cases one gets the very characteristic F1-down, F2-up, F3-down symmetrical pattern. For the /)C)/ cases the transitions are straight. For the mixed vowel cases it seems that the locus that the transitions point to depends on the identity of the first vowel.

For speaker 4:

For the cases when the consonant is bilabial:

(1) Although the bilabial bursts are not spectacular they are very standard in shape and the hump is always in the range 1.3-5 kHz and almost always extends from 1.4-3.7 kHz. Usually the burst is very diffuse and rather flat; however some examples are slightly peaky.

(2) There are several occurrences of burst-cum-formant structures; these almost always occur in /VC V/ cases.

(3) There are no signs of coarticulation in the positions of the burst humps.

(4) There is always at least slight aspiration in all /V CV/ cases and in all unvoiced /VC V/ cases. It is almost always absent in the voiced /VC V/ cases.

(5) There is silence in almost all /V CV/ cases. It also occurs in some

/VC V/ cases (mixed voiced and unvoiced).

(6) Transitions are often not apparent. However the F3 transition is very strong in the /iCi/ cases. Sometimes, unexpectedly, the transition associated with the vowel on the same side of the word boundary as the consonant is missing.

For the cases when the consonant is alveolar:

(1) /t/ bursts tend to be more prominent than /d/ bursts.

(2) This speaker's bursts are very regular in shape. The hump occurs in the range 1.6-5.2 kHz.

(3) Slight coarticulation effects can be seen. In the /iCi/ cases the burst tends to start at 1.8 kHz. It starts at a lower position for the mixed vowel cases and, what is really puzzling, starts at 2 kHz for the two /) C)/ cases.

(4) There are some burst-cum-formant structures in the voiced /VC V/ situations.

(5) There is always aspiration in the /V CV/ cases and in all the unvoiced /VC V/ cases.

(6) Noticeable silence never occurs in the /VC V/ cases and only in a couple of unvoiced /V CV/ cases.

(7) Transitions are most interesting. In all the /iCi/ cases the following transitions are well marked but the preceding transitions are not well marked.

(8) For the /iC)/ cases the preceding F3 transitions and the following

transitions are well marked but the F2 transitions are invariably straight in this case.

(9) For the /)Ci/ cases the transitions are very well marked as they are for the /)C)/ cases. In this latter situation the symmetrical F1-down, F2-up, F3-straight (or slightly down) pattern is well marked.

For the cases when the consonant is velar:

(1) There are many one-peak bursts. Particularly in the /iCi/ cases the hump is diffuse with a peak at approximately 3 kHz in the /k/ cases or two close-together peaks e.g. 2.4 and 3.2 kHz in the /i gi/ case. Similar situations occur in the /)Ci/ cases where peaks are close together (2.1 and 4.1 kHz) or, in one case, with just one peak at 3.0 kHz.

(2) As pointed out coarticulation effects are noticeable in that all the /VCi/ cases have their first peaks high (i.e. >2 kHz). For the /)C)/ cases the first peak is low (0.9-1.4 kHz with second peak at 4.2 kHz). A curious situation occurs for the /iC)/ cases however; the /iC)/ cases have first peaks high (at 2.2 kHz) while the /i C)/ cases have first peaks low (1.6 kHz).

(3) Aspiration occurs in almost all /V CV/ cases and in some /VC V/ cases.

(4) Silence is always associated with unvoiced /V CV/ cases but occurs sporadically in other cases as well.

(5) One striking feature of the transitions is that in the /iCi/ cases the F3 transition is either not present or is very slight.

(6) Transitions for all the /iC)/ cases seem to point to an /i/ locus.

Transitions for the /) Ci/ cases point to an /)/ locus and for the /)C i/ they point to an /i/ locus. Again it appears that the transition locus is associated with the first vowel.

(7) Transitions for the /)C)/ cases are straight.

For speaker 5:

For the cases when the consonant is bilabial:

(1) Bursts are not the standard bilabial shape. They tend to be peaky, in many cases looking somewhat like velar bursts with broad peaks. The general range of the 'hump' is 1.2-3.6 kHz but sometimes it is much wider extending up to about 5.5 kHz. The bursts are not very marked.

(2) In three cases there is no burst. In four cases there is a burst-cum-formant structure. There are no signs of coarticulatory effects.

(3) Aspiration appears a least slightly in all the unvoiced cases. Actually it appears slightly in the unvoiced /VC V/ cases and definitely in all the unvoiced /V CV/ cases. It is missing in all but one voiced case.

(4) Silence appears in most cases; in fact it is only missing in two the two /ibi/ and the two /)bi/ cases.

(5) It should be noted in conjunction with (4) that this speaker often shows inter-word pauses. Also he habitually speaks slowly and distinctly. Transitions are clearly marked in the expected directions. Where there is a fairly long time gap transitions tend to only appear between the vowel and consonant on the same side of the

gap.

(6) Where both sets of transitions are present the following transitions are often more clearly marked than the preceding transitions.

For the cases when the consonant is alveolar:

(1) Bursts are not spectacular. For the /iCi/ cases they tend to be very diffuse and very wide in the range 1-7 kHz and often cover most of the range.

(2) The /iC)/ with the exception of the /id)/ burst which is like the /iCi/ burst tend to be more conventionally alveolar in shape and occur in the range 1.2-4 kHz but the hump is only about 1.4 kHz in width.

(3) The /)Ci/ bursts are a mixture between the two types of bursts described in (1) and (2) above.

(4) Three of the four /)C)/ bursts are rather diffuse velar in shape

(5) At least slight aspiration occurs in all the unvoiced cases and in some voiced cases.

(6) There is no silence in the /VC V/ cases although there is in several /V CV/ cases.

(7) Again the word boundary is often well-marked; at least the drop in waveform amplitude is almost always significant.

(8) Transitions are well-marked in all cases except possibly the /iCi/ cases. This is not surprising however.

For the cases when the consonant is velar:

(1) Bursts are generally the standard velar shape even if they are not unduly spectacular. They show coarticulation effects very well. The /iCi/ bursts are rather diffuse and have peaks at 2.7 and 4.2 kHz.

(2) The /iC)/ bursts are more /)/-ish than /i/-ish with peaks at 1.3 and 3.6 kHz (with the exception of /ig)/ which is a one-peak burst.

(3) The /)Ci/ cases tend more /i/-ish than /)/-ish with peaks at 2.1 and 3.9 kHz. Here the exception is /)ki/ which has /)/-ish peaks at 1.3 and 3.1 kHz.

(4) For the /)C)/ cases the bursts are three-peak in shape with peaks at 0.8, 3.5, and 6.0 kHz. It is interesting to note the presence of this high frequency peak.

(5) Aspiration occurs in all unvoiced cases and slightly in some voiced cases.

(6) Silence only occurs in very few cases although there are big amplitude drops between the vowels.

(7) Transitions are on the whole well-marked although F3 transitions are not so marked as for some speakers. The transitions tend to be strongest with the vowel on the same side of the consonant as the word boundary. Hard to say where the mixed vowel locus is.

For speaker 6:

For the cases when the consonant is bilabial:

(1) Bursts are generally bilabial in shape. The hump occurs in the range

1.1-5.5 kHz, most typically going from 1.5-4.9 kHz. No coarticulation effects are noticeable.

(2) All the voiced /VC V/ cases are burst-cum-formant structures.

(3) Some bursts particularly the /)pi/ bursts are very peaky.

(4) Two bursts, the /i p)/ and the /)p)/ look just like velar bursts.

(5) All the unvoiced cases have at least slight aspiration. None of the voiced /VC V/ has any aspiration but two of the voiced /V CV/ cases have slight aspiration.

(6) At best the silence if it occurs is brief. It tends to occur most commonly in the /V CV/ cases.

(7) Transitions are almost always well-marked - for F3 as well as F2.

For the cases when the consonant is alveolar:

(1) Alveolar bursts are very prominent with massive humps in the range 1.1-6.0 kHz.

(2) All the voiced /VC V/ cases are burst-cum-formant in structure.

(3) The voiceless bursts are generally more massive than the voiced bursts.

(4) In the two /id)/ cases, the two /)d)/ cases, and the /)d i/ case the burst is much narrower than in other cases and is in the range 1.1-3.5 kHz.

(5) No coarticulation effects are noticeable in the bursts.

(6) Aspiration is present in all voiceless cases but in practically no

voiced cases.

(7) Silence occurs in no /VC V/ cases. But it does occur in the voiceless /V CV/ cases.

(8) transitions are all very well marked.

(9) It is noteworthy that in transition from /i/ F2 generally goes slightly up. In these cases F3 is generally strongest.

(10) The /)C)/ transitions take the usual form but are spectacularly marked. Even F4 transitions are very prominent in the /) C)/ case.

For the cases when the consonant is velar:

(1) Well marked two-peak velar bursts. No burst-cum-formant cases.

(2) Evidence of coarticulation. The /iCi/ bursts are at 2.7, 5.1 kHz. All the mixed vowel cases are at 1.6 and 4.7 kHz. This suggests that both vowels have equal effects. All the /)C)/ cases are at 0.7 and 4.5 kHz. The two voiced cases are actually 3 peak bursts with a small peak at 2.6.

(3) As for other speakers the /iCi/ case has the most diffuse peaks. All other cases have really sharp peaks.

(4) There is aspiration in all but one /V CV/ case and in all unvoiced /VC V/ cases. It only appears slightly in one voiced /V CV/ case.

(5) There is silence (at least briefly) in all unvoiced cases and in some voiced /V CV/ cases.

(6) Transitions are well marked with F1 down in almost all cases.

(7) Locus effects suggest that the locus is associated with the previous vowel.

For speaker 7:

For the cases when the consonant is bilabial:

(1) For all but one case the bilabial bursts are all the same shape and essentially overlap. They have a wide-bandwidth peak 1.6-2.4 kHz and then trail off gradually. The general hump goes typically from 1.3-4.8 kHz. There is no sign whatsoever of coarticulation effects.

(2) Aspiration occurs in over half the cases but it is often slight and follows no set pattern.

(3) Silence occurs in all cases. There are often quite large gaps between words.

(4) Transitions are well marked; particularly the following transitions. The only cases where no transitions are seen are in long time gap cases.

For the cases when the consonant is alveolar:

(1) Very marked alveolar bursts indeed.

(2) Three of the six voiced /VC V/ structures are burst-cum-formant.

(3) The /iCi/ case has a wide humped burst going typically from 2.0 to 6.0 kHz. Also it has a peaked structure of a narrow peak and then a wide hump.

(4) For all the mixed vowel cases the hump is typically 1.6-6.5. It tends

to be slightly wider in the /)Ci/ case than in the /iC)/ case.

(5) The /)C)/ cases are generally massive bursts, somewhat narrower in hump than other bursts. They typically extend from 1.5 to 3.9 kHz.

(6) The voiceless bursts are overall more massive than the voiced bursts.

(7) Aspiration occurs in all the /V CV/ cases and all the voiceless /VC V/ cases.

(8) Silence occurs in all the /VC V/ cases and only occurs once in the /VC V/ cases.

(9) Transitions are very well marked as expected except where the time gap is very large.

(10) It is interesting to note that F3 goes down in the /)C)/ cases.

For the cases when the consonant is velar:

(1) Bursts well marked. The /iCi/ cases have the peaks at 3.4 and 4.6 kHz approximately. In many ways it is more like a high peaky region from 2.0-5 kHz.

(2) The mixed vowel cases show the influence of both vowels with the second vowel having the strongest effect. Thus for /iC)/ case peaks are at 1.5 and 4.4 kHz. (Slight exception in /ik)/ where first peak is 2.1 kHz.)

(3) For the /)Ci/ case the peaks are typically at 2.5 and 4.2 kHz. The very odd exception here is /) ki/ (1.6 and 4.4 kHz).

(4) The /)C)/ case has the lowest peaks all at 0.8 and 4.0 kHz. The two unvoiced cases are multiple peak bursts, but still clearly velar.

(5) Aspiration in all /V CV/ cases and all voiceless /V CV/ cases.

(6) Silence at least slight in almost all cases (exception /igi/).

(7) Transitions generally well marked including F3. Generally they point to locus of the preceding vowel. Exceptions occur in conjunction with long time gaps.

For speaker 8:

For the cases when the consonant is bilabial:

(1) Bursts are well marked and the expected shape with the hump in the range 0.6-6.0 kHz - most typically occurring in the range 1.0-5.4 kHz.

(2) All the voiced /VC V/ cases are burst-cum-formant structures.

(3) There is no evidence of coarticulation on the position of the hump.

(4) There is aspiration in all the /V CV/ cases and in three voiceless /VC V/ cases.

(5) There is silence in all /V CV/ cases and in all voiceless /VC V/ cases.

(6) The transitions are very well marked indeed, including the F3 transitions. The only exception is when the time gap is very large.

For the cases when the consonant is alveolar:

(1) Bursts are not terribly strong except the two /) C)/ cases. Yet most are still fairly recognizably alveolar.

(2) Most bursts are wide rather than excessively high and are in the range

1.0-6.6 kHz, typically going from 1.6-5.5 kHz.

(3) Three of the four voiced /VC V/ bursts are burst-cum-formant structures.

(4) All /V CV/ cases are aspiration as are all voiceless /VC V/ cases but no voiced /VC V/ cases.

(5) There is silence in al but one /V CV case and in all but one voiceless ?VC V/ case, but not in the voiced version of the last case.

(6) Transitions are well marked in the expected directions except for the two /)C i/ cases where F3f goes down and the /)d)/ case when F2 stays practically straight the whole time.

For the cases when the consonant is velar:

(1) Bursts typically velar. The /iCi/ bursts can be one peak or like a high alveolar burst. The 2 peaks are approximately at 2.8 and 5.0 kHz.

(2) Coarticulation effects are strong with the /iC)/ case peaks being at 1.4 and 4.9 (except for /ik)/, with peaks at 2.0 and 5.3 kHz, but then there is a long time gap between the /k/ and the /)/). The /iC)/ burst peaks are at 2.3 and 4.7 kHz (again except for a time gap exception /)ki/).

(3) There are no bursts in the /)C)/ cases but the peaks for the /) C)/ cases are at 1.0 and 4.6 kHz.

(4) Aspiration occurs in all /V CV/ cases and in most /VC V/ cases.

(5) Silence in all but one /V CV/ case and in all the voiceless /VC V/

cases.

(6) Transitions are generally well marked with a slight tendency for the locus to be associated with the vowel on the same side of the word as the consonant.

(7) F3 transitions are particularly strong.

For speaker 9:

For the cases when the consonant is bilabial:

(1) Bursts are standard bilabial bursts, very diffuse and in this case vary widely going from 1.0-7.0 kHz and most typically from 1.5-5.5 kHz.

(2) There are no bursts in three of the /VC V/ cases, viz: /ibi/, /ib)/ and /)p)/.

(3) There is no sign of coarticulation in the burst.

(4) Aspiration is not very marked. It is distinctly noticeable in all the voiceless /V CV/ cases however.

(5) Silence occurs in all voiceless cases, and in three vowel cases.

(6) Transitions are invariably present. F3 is always very well marked.

(7) It is curious that in /ib i/, /ip i/ and /ib)/ cases the F2 following transition is very steep indeed.

For the cases when the consonant is alveolar:

(1) Bursts are generally reasonable if not spectacular alveolar bursts.

(2) Bursts show coarticulatory influence in that bursts associated with /i/ tend to be highest in the high frequency range and the hump in this case goes typically for 2.6-6.0 kHz. These bursts tend to be a bit diffuse.

(3) Bursts associated with /)/ are more compact that bursts associated with /i/. Generally they are fairly high over a narrow frequency range e.g. 1.4-4.6 kHz.

(4) Both vowels seem to affect the position of the burst.

(5) Three of the voiced /VC V/ cases show traces of burst-cum-formant structure.

(6) Aspiration occurs in all /V CV/ cases and is very marked. It occurs somewhat randomly in the /VC V/ cases but is present in all but one velar /VC V/ case.

(7) There is no silence in any voiced case but it occurs in all but two /VC V/ instances of the voiceless cases.

(8) Transitions are well marked and very considerably present. Transition directions are internally very consistent for each VCV case. In /i/ cases F2 definitely goes down.

(9) F4 transitions are apparent in some cases running parallel to the F3 transitions.

(10) In all the /iCi/ cases F3 goes down strongly whereas in all the /iC)/ cases F3 stays straight.

(11) Voiceless bursts tend to be higher in frequency position than voiced

bursts.

For the cases when the constant is alveolar:

(1) The /iCi/ bursts tend to be one peak (at 2.9 kHz) with a second smaller peak sometimes visible, the exception is /i ki/ with 2 clear peaks at 2.6 and 5.2.

(2). The /)Ci/ cases have the first burst peak in the range 1.7-2.7 kHz and second peak in the range 4.3-5.1 kHz.

(3) The /iC)/ case has the first peak in the range 1.1-1.6 kHz and the second peak at 4.4 kHz.

(4) The two /VC V/ cases have no burst but the two /) C)/ bursts have peaks at 1.2 and 4.6.

(5) There is aspiration in all /V CV/ cases and in most voiceless /VC V/ cases.

(6) Silence appears randomly in half the cases.

(7) Transitions are quite well marked although F2 does not go up in the /ig i/ case.

(8) There is a tendency for the locus to be associated with the first vowel.

For speaker 10:

For the cases when the consonant is bilabial:

(1) Bursts tend to be long flat bursts (i.e. extremely diffuse) and ranging from 0-6 kHz.

(2) Most of the bursts show burst-cum-formant structure. There are two cases where no burst is produced - /) b)/ and /)p)/.

(3) Some of the /)/ bursts seem extremely low (0-1.5 kHz). It is hard to tell in these cases whether it is a burst-cum-formant structure or just burst.

(4) There is some evidence of coarticulation - /i/ bursts start at 1.7-2 kHz whereas /)/ bursts are as described in (3) or at least they start at 1.2 kHz.

(5) There is very little aspiration, certainly none in the voiced cases. when it occurs as it does in five voiceless cases it takes the form of a strange periodic structure rather than noise.

(6) Silence is present in all but three cases.

(7) Transitions are well marked although the following transitions are usually stronger than the preceding transitions in the /VC V/ cases.

(8) Note this speaker very often produced /i/ as [)i] particularly when /i/ in the penultimate position of a word.

For the cases when the consonant is alveolar:

(1) Bursts are all strongly marked in the typical alveolar range.

(2) Some of the voiced /VC V/ cases show a burst-cum-formant structure.

(3) Coarticulation effects are noticeable in that the hump for the /i/ cases is generally wider than for the /)/ cases. Both vowels affect the position of the hump. Thus for /iCi/ the hump is 2.2-4.5 kHz for the /VC V/ cases and 3-6 kHz for the /V CV/ cases. For /)Ci/ the hump

is 1.8-4.7 kHz for the /VC V/cases and 2.5-5.7 kHz for the /V CV/ cases. For /iC)/ the hump is 1.8-4.3 kHz for all cases. For /)C)/ the hump is 1.2-2.5 kHz for the /VC V/ cases and 1.8-3.8 kHz for the /V CV/ cases. From this it can be seen that the burst in the /V CV/ cases is in general higher and wider in frequency that for the burst in the /VC V/ cases.

(4) Aspiration occurs in all voiceless /V CV/ cases and in a few other cases.

(5) Silence occurs in all voiceless cases and in two voiced cases.

(6) There are no transitions associated with /i/ in any /iC V/ cases. There is a strongly marked F2f transitions in the /iC V/ case however.

(7) The F2 transitions in the /)Ci/ cases are all well marked.

(8) F1, F2 and F3 transitions are all well marked in the /)C)/ case.

For the cases when the consonant is velar:

(1) Well marked velar bursts. In the /i)i/ case the burst is basically one-peak at 3 kHz with sometimes a smaller peak at 5 kHz.

(2) In the /)Ci/ case the peaks are at 1.9 and 4.7 kHz for the /VC V/ cases and 2.3 and 4.5 kHz for the /V CV/ case.

(3) In the /iC)/ cases, the first peak is in the range 1.2-1.9 kHz and the second peak in the range 4.6-5.1 khz.

(4) All the /)C)/ peaks are at 0.9 and 5.1 khz.

(5) Thus it seems that both vowels have an effect on the position of the peaks and that the position of the juncture plays a part in

determining the amount of influence of each peak.

(6) There is aspiration in all but one voiceless case and (randomly) in half the voiced cases.

(7) There is silence in all voiceless cases and (randomly) in half the voiced cases.

(8) The transitions are all well marked in the appropriate direction and the loci tend to be associated with the first vowel although in one case, /i g)/, it is associated with the second vowel.